Quasi real-time model for security of water distribution network

Hervé Ung

université de BORDEAUX

THÈSE

PRÉSENTÉE À

L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE DE MATHÉMATIQUES ET INFORMATIQUES

Par Monsieur Hervé UNG

POUR OBTENIR LE GRADE DE

DOCTEUR

SPÉCIALITÉ : MATHÉMATIQUES APPLIQUÉES ET CALCUL SCIENTIFIQUE

Titre : **Quasi Real-Time Modeling for Security of Water Distribution Networks**

Soutenue le 05 Février 2016

Après avis de MM.

Thierry Horsin	Professeur des Universités, CNAM Paris	Rapporteur
Serge Huberson	Professeur des Universités, Université de Poitiers	Rapporteur

Devant la Commission d'examen formée de :

Mejdi Azaiez	Professeur des Universités, Bordeaux INP	Président du jury
Jochen Deuerlein	Docteur Ingénieur, 3S Consult	Examinateur
Cédric Féliers	Docteur Ingénieur, VEDIF	Invité
Denis Gilbert	Docteur Ingénieur, IRSTEA	Invité
Thierry Horsin	Professeur des Universités, CNAM Paris	Rapporteur
Serge Huberson	Professeur des Universités, Université de Poitiers	Rapporteur
Angelo Iollo	Professeur des Universités, Université de Bordeaux	Co-directeur
Pierre Mandel	Docteur Ingénieur, Veolia Recherche et Innovation	Examinateur
Iraj Mortazavi	Professeur des Universités, CNAM Paris	Co-directeur
Olivier Piller	Chargé de recherche, IRSTEA	Co-directeur

Remerciements

Je voudrais d'abord remercier toute l'équipe Irstea (Cemagref lorsque j'ai commencé) de m'avoir si bien accueilli dans ses locaux. Je remercie aussi l'Institut de Mathématiques de Bordeaux et tous mes collègues et nombreux co-bureaux.

I would like to thank all my partners of the SMaRT-OnlineWDN project, and particularly Dr. Delmouly and Dr. Krug who were the ANR and VDI reviewers and have supported it. It has been a project full of conviviality and I have had the opportunity to meet some people of great qualities.

Je voudrais accorder un grand merci à mes deux rapporteurs T. Horsin et S. Huberson pour leurs conseils et suggestions qui ont contribué à l'amélioration de ce manuscrit.

Je veux remercier O. Piller pour m'avoir soutenu tout au long de cette thèse, ainsi que pour m'avoir apporté ses connaissances et son expérience internationale.

Merci à I. Mortazavi de m'avoir donné l'opportunité de faire cette thèse et de m'avoir fourni ses conseils pleins de sagesse.

J'aimerais remercier A. Iollo pour ses conseils avisés à chaque fois que j'en ai eu besoin.

Merci bien sûr à D. Gilbert, pour son soutien constant qui m'a permis de tenir le cap ainsi que son expertise technique qui a apporté un vrai plus au cours de ma thèse.

Mes remerciements vont aux autres membres du jury : M. Azaiez pour avoir accepté d'être président de jury, J. Deuerlein pour sa jovialité et son expérience industrielle, C. Féliers pour sa gentillesse et son expérience du terrain, P. Mandel pour son amabilité et son expérience technique.

Un grand merci à tous mes collègues d'Irstea, à l'équipe GPIE et à toutes les personnes qui ont été de passage chez nous. Un coucou à tous les gens du club café qui vont me manquer, à tous ces bons moments autour d'un bon petit verre chaud en espérant qu'ils dureront longtemps.

Finalement je n'oublierai pas ma famille, ma copine et mes amis qui m'ont toujours soutenu même dans les moments difficiles et je les en remercie grandement.

Quasi Real-Time Modeling for Security of Water Distribution Networks

Summary

The aim of this thesis is to model the propagation of a contaminant inside a water distribution network equipped with real time sensors. There are three research directions: the solving of the transport equations, the source identification and the sensor placement.

Classical model for transport of a chemical product in a water distribution network is using 1D-advection-reaction equations with the hypothesis of perfect mixing at junctions. It is proposed to improve the predictions by adding a model of imperfect mixing at double T-junctions and by considering dispersion effect in pipes which takes into account a 3-D velocity profile. The first enhancement is created with the help of a design of experiment based on the Delaunay triangulation, CFD simulations and the interpolation method Kriging. The second one uses the adjoint formulation of the transport equations applied with an algorithm of particle backtracking and a random walk, which models the radial diffusion in the cross-section of a pipe.

The source identification problem consists in finding the contamination origin, its injection time and its duration from positive and negative responses given by the sensors. The solution to this inverse problem is computed by solving the adjoint transport equations with a backtracking formulation. The method gives a list of potential sources and the ranking of those more likely to be the real sources of contamination. It is function of how much, in percentage, they can explain the positive responses of the sensors.

The sensor placement is chosen in order to maximize the ranking of the real source of contamination among the potential sources. Two solutions are proposed. The first one uses a greedy algorithm combined with a Monte Carlo method. The second one uses a local search method on graphs.

Finally the methods are applied to a real test case in the following order: the sensor placement, the source identification and the estimation of the contamination propagation.

key words: Water distribution network – Source identification – Backtracking – CFD - Imperfect mixing – Double T-junction – Dispersion – Random walk – Monte Carlo – Greedy algorithm – SMaRT-Online^{WDN}

Research Unit: Environnement, Territoire et Infrastructure (ETBX), Irstea, 50 avenue de Verdun, 33612, Cestas (France)

Modèle quasi-temps réel pour la sécurité des réseaux d'alimentation en eau potable

Résumé

Le but de cette thèse est de modéliser la propagation d'un contaminant au sein d'un réseau d'eau potable muni de capteurs temps réel. Elle comporte les trois axes de développement suivant: la résolution des équations de transport, celle du problème d'identification des sources de contamination et le placement des capteurs.

Le transport d'un produit chimique est modélisé dans un réseau d'eau potable par l'équation de transport réaction 1-D avec l'hypothèse de mélange parfait aux nœuds. Il est proposé d'améliorer sa prédiction par l'ajout d'un modèle de mélange imparfait aux jonctions double T et d'un modèle de dispersion prenant en compte un profil de vitesse 3-D et la diffusion radiale. Le premier modèle est créé à l'aide d'un plan d'expériences avec triangulation de Delaunay, de simulations CFD, et de la méthode d'interpolation krigeage. Le second utilise les équations adjointes du problème de transport avec l'ajout de particules évoluant à l'aide d'une marche aléatoire, cette dernière modélisant la diffusion radiale dans la surface droite du tuyau.

Le problème d'identification des sources consiste, à l'aide de réponses positives ou négatives à la contamination des nœuds capteurs, à trouver l'origine, le temps d'injection et la durée de la contamination. La résolution de ce problème inverse est faite par la résolution des équations de transport adjointes par formulation backtracking. La méthode donne la liste des sources potentielles ainsi que le classement de celles-ci selon leur probabilité d'être la vraie source de contamination. Elle s'exprime en fonction de combien, en pourcentage, cette source potentielle peut expliquer les réponses positives aux capteurs.

Le placement des capteurs est optimisé pour l'identification des sources. L'objectif est la maximisation du potentiel de détection de la véritable source de contamination. Deux résolutions sont testées. La première utilise un algorithme glouton combiné à une méthode de Monte Carlo. La seconde utilise une méthode de recherche locale sur graphe.

Finalement les méthodes sont appliquées à un cas test réel avec dans l'ordre : le placement des capteurs, l'identification de la source de contamination et l'estimation de sa propagation.

Mots clés: Réseau d'eau potable – Identification des sources – Backtracking – CFD – Mélange imparfait – Double T – Dispersion - Marche aléatoire – SMaRT-Online^{WDN}

Unité de recherche : Environnement, Territoire et Infrastructure (ETBX), Irstea, 50 avenue de Verdun, 33612, Cestas (France)

Résumé de la thèse : Quasi Real-Time Modeling for Security of Water Distribution Networks

Auteur: Hervé Ung

La contamination d'un réseau de distribution d'eau potable, délibérée ou non, peut entrainer de sérieuses conséquences sur la santé et la sécurité des consommateurs. Pour prévenir cela et pour mieux comprendre l'état des réseaux, les opérateurs du projet SMaRT-Online^{WDN} ont décidé d'installer des nouveaux capteurs temps réel qui peuvent mesurer la qualité de l'eau. Ils mesurent les paramètres qualités tels que la conductivité, le pH, la turbidité, le taux de chlore libre ou encore la température. En couplant les données mesurées avec un algorithme de génération d'alarme, il est possible de détecter une contamination à l'intérieur du réseau. En effet, même si les capteurs ne mesurent pas la concentration d'un produit spécifique, le produit contaminant réagit généralement avec le milieu environnant et donc modifiera au moins un ou deux des paramètres qualités mesurés.

Deux réponses de contamination peuvent êtres alors obtenues: soit positive soit négative. En couplant ces réponses au temps de détection, il est possible de retracer dans le temps avec un algorithme de marche rétrograde et de construire la liste des nœuds potentiellement sources de contamination. Une source potentielle est représentée par un lieu (un nœud du réseau), un temps de début de contamination et d'une durée de contamination. La probabilité pour chaque nœud potentiel d'être la véritable source de contamination peut être calculée. Cette information permet aux opérateurs de réagir de deux manières. La première est de stopper l'injection de contaminant dans le réseau en trouvant sa source. La deuxième est d'empêcher la propagation du contaminant dans le réseau en utilisant différentes vannes et poteaux d'incendie pour protéger différents districts et diriger l'eau contaminé dans des lieux sécurisés où elle pourra être extraite. Pour faire cela, un modèle de propagation fiable est nécessaire pour déterminer les actions à prendre de façon à limiter les dommages humains et matériels.

Cette thèse a été finance par le projet franco-allemand appelé SMaRT-Online^{WDN}, il réunit 8 partenaires : trois opérateurs de réseau, Berliner Wasserbetriebe (BWB), le service d'eau de la CUS, Veolia Eau d'Ile-de-France (VEDIF); 4 instituts de recherches, ENGEES UMR, Fraunhofer IOSB, Irstea et DVGW TZW (Dresden); et 2 industriels, 3S Consult GmbH et Veolia Environnement (VERI). Le projet a pour but le développement d'un logiciel de sécurité des réseaux d'eau potable dans le cas d'une contamination, accidentelle ou intentionnel, du réseau de distribution d'eau potable. Les différents aspects étudié sont : le placement des capteurs, la génération d'alarme, l'identification des sources de contamination, l'amélioration du modèle de transport, la prévision de la demande, la conception du logiciel et les enquêtes sur la réaction des consommateurs français et allemands à une contamination du réseau.

L'objective de cette thèse se porte sur le développement de trois modèles : l'amélioration du modèle de transport (mélange imparfait aux jonctions en double Tés et dispersions dans les

tuyaux), l'identification des sources et le placement des capteurs. La thèse aborde ces thèmes sur six chapitres. Le premier décrit les modèles actuels d'hydraulique et de transport ainsi que des solutions apportées à certains problèmes dans les modèles de sécurité des réseaux distribution d'eau potable. Le second et troisième chapitre traite du modèle de transport avec le mélange imparfait aux jonctions en double Tés et la dispersion dans les tuyaux. Le quatrième explique le problème d'identification des sources et donne une solution pour construire la liste des nœuds potentiels classé suivant le nœud ayant le plus de chance d'être la véritable source de contamination. Un critère est aussi donné pour évaluer un tel modèle d'identification de source. Le cinquième chapitre utilise le critère défini précédemment pour le placement de capteurs adapté au mieux à l'identification de sources. Deux méthodes ont été développées, une méthode gloutonne et une formulation de recherche locale sur graphe. Finalement, le dernier chapitre est une étude de cas utilisant les différentes méthodes développées ci-avant sur un grand réseau.

Les réseaux de distributions d'eau potable sont composés de nœud jonctions, réservoirs, ressources et de tuyaux. La dynamique de l'eau à l'intérieur de ces éléments est fixée par la demande aux nœuds de consommations, ainsi que par la hauteur d'eau aux nœuds ressources et par la rugosité dans les tuyaux. Les nœuds ressources représentent une source d'eau intarissable, avec une hauteur d'eau fixe, et l'eau généralement pompée permet la satisfaction de la demande. Les nœuds réservoirs contiennent un certain volume d'eau qui varie le long de la journée. Et les nœuds jonctions peuvent être à la fois intersections de tuyaux ou bien nœud consommateur.

Dans les model de distribution d'eau potable, le transport est simplifié et est modélisé par l'équation d'advection réaction 1-D dans les tuyaux et une condition de parfait mélange aux jonctions. Il est montré dans la littérature que les deux simplifications peuvent introduire de grandes erreurs dans les calculs. Ho et al. (2008) et Choi et al. (2008) ont montré que le mélange imparfait apparaissait dans le cas de croix et de jonctions en double Tés. Dans cette thèse, des simulations CFD de mélange aux jonctions en double Tés ont été réalisées pour remplir un tableau de données sur lequel une interpolation est utilisée pour définir une loi 1-D pour le mélange imparfait aux jonctions en double Tés. Les logiciels Code Saturne et ANSYS Fluent ont été utilisé pour faire les calculs, sur grille, respectivement en laminaire et en turbulent. Un modèle d'interpolation Krigeage est développé et optimisé dans le cas d'un plan d'expérience non structuré. La triangulation de Delaunay est ensuite couplée à l'erreur d'interpolation pour définir le prochain point de calcul, celui où l'erreur est la plus grande. Les calculs sont ensuite comparés aux expériences menées dans le laboratoire à Dresde et une loi 1-D est créé et implémenté dans le modèle de transport.

De plus, Taylor (1993) et d'autres ont montrés une différence dans le temps d'arrivé d'un agent chimique si on prend en compte le profile de vitesse 3-D et la diffusion dans la section droite. Un coefficient virtuel de dispersion est généralement utilisé pour modéliser un tel phénomène. Dans cette thèse, les équations 3-D de l'advection ont été directement utilisé avec l'aide d'un modèle de marche rétrograde couplé à un model de marche aléatoire modélisant la diffusion radiale. Il est montré qu'un tel modèle marche aussi pour des petits coefficients de

diffusion qui ne peuvent pas être modélisés avec l'ajout d'un terme de diffusion à l'équation d'advection 1-D utilisé habituellement. Une étude d'adimentionnalisation des équations d'advection diffusion 3-D est effectuée montrant qu'on ne peut négliger la diffusion radiale dans la section droite. Une formulation de marche rétrograde des particules est utilisée pour modéliser l'advection comportant un profil de vitesse et une marche aléatoire permet le calcul de la diffusion radiale. Les résultats montrent un bon comportement en régime laminaire avec les expériences. Moins proche, on observe une dispersion toujours présente en régime turbulent.

L'identification des sources consiste à déterminer le lieu d'injection des sources de contaminations, leurs temps d'injections et leurs durées d'injection. On utilise pour cela les réponses positives et négatives aux capteurs. Une fois que les nœuds potentiels de contamination sont déterminés, les équations améliorées du transport permettent le calcul de la propagation de la contamination présente et future. Des multiples travaux ont été réalisés à ce sujet. La thèse complète les travaux précédant de Shang et al. (2002) et de Propato et al. (2010) sur la marche rétrograde de particule et la matrice de contamination. Sur le même procédé, un modèle de marche rétrograde est utilisé pour créer la matrice d'entrée/sortie des contaminations à partir des réponses aux capteurs. La méthode a été modifiée afin de pouvoir effectuer le calcul sur le graphe d'un grand réseau. La matrice d'entrée/sortie des contaminations donne et classe les éléments de la liste des nœuds de contamination. Elle permet aussi de repérer les cas de contaminations multiples. De plus, des critères sont définis afin d'évaluer la méthode d'identification des sources sur une configuration spécifique de placement de capteurs. Une étude par la méthode de Monte Carlo permet de regrouper les scénarios suivant trois catégories : non détecté, non précis et spécifique. Dans les deux premiers cas, la source de contamination ne fait pas partie de la liste des nœuds potentiels, soit à cause du placement des capteurs ou bien de la méthode d'identification des sources. Il n'y a seulement que dans le dernier groupe que la source est trouvée et classée, sa position dans le classement déterminant la capacité de spécification de la méthode.

Le placement des capteurs est un problème crucial pour les manageurs de réseaux de distribution d'eau potable. En effet, leurs caractéristiques, leurs nombres et leurs localisations influencent grandement la capacité de surveillance du réseau ainsi que l'efficacité des méthodes d'optimisation sur lequel elles s'appliquent. Par exemple, l'efficacité de l'identification des sources dépend du nombre et de la localisation de ces capteurs. Un capteur à entrées multiples qui peut mesurer la conductivité, la température et le taux de chlore libre est cher mais recueille beaucoup de données importantes. Un capteur qui ne mesure qu'un seul de ces paramètre sera lui limité en donné, mais il ne coutera peu et un nombre important de capteur pourra alors servir pour un couvrement important du réseau. Beaucoup de travaux se sont attaché à l'emplacement des capteurs, les opérateurs choisiront surement les capteurs chers, car le coût d'installation doit aussi être pris en compte en plus du coût du capteur. De plus, le coût de système de télé relève, type SCADA, ainsi que le salaire de l'équipe de maintenance et les coûts associés (alimentation, réparation,…) sont aussi à être pris en compte. Beaucoup de méthodes de placement de capteurs

ont été crées pour la bataille des capteurs de réseaux de distribution d'eau (Battle of the Water Sensor Network) détaillé dans le papier de Ostfeld et al. (2008). La pluparts utilisent des critères généraux tel que la vraisemblance de détection ou bien le temps moyen à la détection. Cette thèse utilise le critère d'identification des sources développé dans le chapitre 4 afin de l'adapter à l'emplacement optimal des capteurs pour trouver et mieux classer la véritable source de contamination. Deux méthodes sont proposées, une méthode gloutonne qui consiste à rajouter à la liste l'emplacement de capteur qui permet le gain maximal sur le critère d'identification des sources. Elle donne de très bons résultats mais est très lente. L'autre méthode est celle de recherche locale sur graphe qui consiste à partir d'un emplacement prédéfini de capteur et à trouver la solution proche qui maximise le critère. Elle donne de bon résultat mais est plus rapide.

Finalement, un cas d'étude sur un modèle de grand réseau est présenté, les quatre méthodes présentées ci-avant sont appliquées dans l'ordre suivant : placement de capteur, identification de la source de contamination et estimation de la propagation présente et future. La méthode gloutonne consomme trop de ressources pour être utilisée dans ce cas. La méthode de recherche locale sur graphe donne de bon résultat avec une initialisation avec un critère de maximisation du temps moyen à la détection. L'utilisation du modèle de transport avec comme condition aux limites les résultats de l'identification des sources permet l'estimation de la propagation de la contamination. Plus le temps d'observation est grand, plus le nombre de sources potentiels augmente, mais plus l'estimation est précise. Finalement, l'influence du mélange imparfait et de la dispersion est estimée. Si dans ce cas l'impact du mélange imparfait peut être négligé, l'effet de la dispersion sur les résultats doit être pris en compte.

Remercions les trois instituts suivants pour nous avoir laissé utiliser leur grille de calcul pour réaliser les simulations :

- Mésocentre de Calcul Intensif Aquitain (MCIA, Avakas)
- Centre Informatique National de l'Enseignement Supérieur (CINES, Jade)
- Irstea Clermont

Cette thèse fait partie du projet *SMaRT-Online*^{WDN} et est financée par le German Federal Ministry of Education and Research (BMBF; project: 13N12180) et par l'Agence Nationale de la Recherche française (ANR; project: ANR-11-SECU-006).

CONTENTS

ILLUSTRATIONS

TABLES

NOTATIONS

A: incidence matrix

q, Q: flow rate in L/s

d: nodal demand in L/s

h: head in m of water

ξ: headloss function in m of water

L: pipe length in m

D: pipe diameter in mm

R: pipe radius in mm

CHW: Hazen Williams coefficient

λ: Darcy friction factor

k: roughness coefficient

u: velocity in m/s

u_0: mean velocity in m/s

C: concentration in mg/L

K: reaction coefficient in 1/h

α: reaction order

r: radius parameter in m, or Hazen Williams resistance

E^*: Taylor virtual diffusion coefficient

D_m: molecular diffusion coefficient in m^2/s

Re: Reynolds number

Sc: Schmidt number

CFD: computational fluid dynamics

DNS: direct numerical simulation

LES: large eddy simulation

1-D: one dimensional

2-D: two dimensional

3-D: three dimensional

Z0: mass flow percentage going out at outlet 1

Z: mass flow percentage going out at outlet 1 minus the perfect mixing result

XX/YY: XX% of the inlet flow comes from the inlet 1 and YY% of the outlet flow goes into the outlet 1 of the double T-junction (see Figure 6).

BT: backtracking time

OT: observation time

ATDOP: average time to detection optimal placement

DLOP: detection likelihood optimal placement

GOP4: greedy optimal placement, backtracking time = 4h

GOP24: greedy optimal placement, backtracking time = 24h

LSATDOP4: local search optimal placement, ATDOP initial placement, backtracking time = 4h

LSATDOP24: local search optimal placement, ATDOP initial placement, backtracking time = 24h

LSDLOP4: local search optimal placement, DL initial placement, backtracking time = 4h

LSDLOP24: local search optimal placement, DL initial placement, backtracking time = 2h

GENERAL INTRODUCTION

The contamination of a water distribution network, deliberate or not, can induce serious consequences on consumer health and security. To prevent that and to better understand networks, SMaRT-Online^{WDN} operators have decided to install real time sensors to measure the water quality. They measure quality parameters such as conductivity, pH, turbidity, free chlorine and temperature. By coupling the data measured with an alarm generation algorithm, it is possible to detect a contamination inside a network. Indeed, even if the sensors do not measure the concentration of a specific product, a contamination product generally reacts with its surrounding and therefore will modify at least one or two measured water quality parameters.

Then two contamination responses can be obtained: either positive or negative. By coupling it to the time of detection, it is possible to go back in time with a backtracking algorithm and find the list of the potential sources of contamination. One potential source is represented by a location (a node of the network), a starting time of contamination and a duration of contamination. The probability of each potential node of being the true source of contamination can be computed. That information allows operators to react in two ways. The first one is to stop the injection of the contamination inside the network when the source location is found. The second one is to mitigate the propagation of the contaminant in the network by using different valves and fire hydrants to protect different areas and direct the contaminated water ways in a secured area to extract it. For that purpose, a reliable propagation model is then necessary to determine the actions to be taken in order to limit human and material damages.

This thesis has been financed by the Franco-German project called SMaRT-Online^{WDN}, gathering 8 partners: three network operators, Berliner Wasserbetriebe (BWB), CUS water service, Veolia Eau d'Ile-de-France (VEDIF); 4 research institutes, ENGEES UMR, Fraunhofer IOSB, Irstea and DVGW TZW (Dresden); and 2 industrials, 3S Consult GmbH and Veolia Environnement (VERI). The project aim is the development of a tool for the security of water distribution networks in case of accidental or intentional contamination. The different aspects that have been investigated are: the sensor placement, the alarm generation, the contamination source identification, the transport model enhancing, the demand forecast, the software conception and the surveys on French and German customer reactions to a contamination in the network.

The objective of this thesis is to develop computer models focusing on three aspects: the enhancement of the transport model (imperfect mixing at double T-junctions and dispersion in pipes), the source identification and the sensor placement. In this regard, this thesis is divided in six chapters. The first one details the current hydraulic and transport models as well as solutions to issues in water distribution network security models. The second and third chapters focus on the transport model on both imperfect mixing at double T-junctions and dispersion in pipes. The fourth one explains the source identification problem and gives a solution to find and rank a list

of potential sources. It also defines criteria in order to evaluate such source identification model. The fifth chapter uses precedent defined criteria to adapt the sensor placement problem to source identification. Two methods have been developed, a greedy and a local search on graph formulations. Finally, the last chapter is a case study that uses the different developed models on a large network.

Water distribution networks are made of: junctions, tank and resource nodes as well as pipes. Water dynamics inside these elements are determined by demands at consumption nodes as well as the head at resource nodes and roughness in pipes. The resource nodes represent an infinite-capacity of water, with a fixed head, and water is generally pumped to satisfy the demand. The tank nodes are containers that have a water level that can vary along the day. Junction nodes are both pipe intersections and consumer nodes.

In the water distribution model, the transport is simplified to the 1-D advection reaction equation in the pipes coupled with perfect mixing at the junctions. It is shown in the literature that both simplifications may induce large errors in the calculation. Ho et al. (2008) and Choi et al. (2008) have demonstrated that imperfect mixing occurred at crosses and double T-junctions. In this thesis, CFD simulations of mixing at double T-junctions have been used to complete a data table upon which an interpolation is used to define a 1-D law for imperfect mixing at double T-junctions.

Also, Taylor (1953) and others have shown a difference in the time of arrival of the chemical agent depending on the velocity profile and the diffusion in the cross section. A virtual dispersion coefficient is generally used to model such phenomena. In this thesis, the 3-D equations have been directly computed with the help of a backtracking model coupled with a random walk scheme to model radial diffusion. It is shown that such model also works for low diffusion coefficients that can't be directly modeled with a diffusion term added to the 1-D transport equation.

The contamination source identification consists in determining the source locations, their injection times and durations. This is done from positive and negative responses at sensors. Once potential contaminations are determined, the enhanced transport equations can then be used to calculate the present and future contamination extent. Multiple works have been done on the subject. This thesis completes previous works by Shang et al. (2002) and Propato et al. (2010) on particle backtracking and contamination matrix. Similarly, a backtracking model is used to create the input/output matrix of contamination from the sensor responses. The method was changed to permit working with large network graph. The IO matrix can both give and rank the potential list of contamination nodes. It can also help to discriminate multi-contamination. Moreover, criteria are defined to evaluate the source identification on a specific sensor placement configuration.

The sensor placement is a crucial issue for water distribution network managers. Indeed, their characteristics, their numbers and their locations influence greatly their networks monitoring

capacity as well as the optimization methods efficiency. For instance, source identification effectiveness depends on the sensors design. A multi-probe sensor that can measure conductivity, temperature and free chlorine is expensive but can bring important information. A sensor that can only measure one of those parameters will be limited in data. However, it will be cheaper and a larger amount of sensors can thus be bought to have a large coverage of the network. Lots of works have been done on the sensor placement design, it appears that the operator will most likely choose to go for expensive sensors because the installation cost also needs to be taken into account no matter the cost of sensors. Besides, the financial balance also needs to include the cost of any SCADA like system, which manages the database, as well as the operating staff salary, and maintenance and operation cost (energy, repair, ...). Multiple methods in optimal sensor placement have been developed in Ostfeld et al. (2008) for the Battle of the Water Sensor Network. Most of them have used general criteria such as detection likelihood and average time to detection. This thesis uses the source identification criteria developed in Chapter 4 to adapt sensor placement to source identification maximization. Two methods are proposed, a greedy formulation that gives very good results but is very slow and a local search on graphs formulation that gives good results but is faster.

Finally, a test case on a large network is presented, applying the four precedent methods that have been developed, in this order: sensor placement, contamination source identification and estimation of present and future propagation. Imperfect mixing and dispersion influence on the model are also described.

CHAPTER I

Water Distribution Network Model for Security

A water distribution network is composed of resources, tanks and consumers as well as pipes, junctions and equipments such as valves and pumps. The usual way to model such networks is through a graph. The incidence matrix defined later is used to formulate the hydraulic problem as defined by Todiny et al. (1987). The current model of drinking water supposes a pressure load network so that the water covers the whole cross section. The models also assume satisfaction of demands, i.e. all consumption nodes get the water they need. The network is represented by a graph composed of nodes and pipes. The first ones are characterized by resources and tanks as well as junctions and demand nodes. The pipes are the links between the nodes and transport the water from sources to consumers. Finally, the equipments are installations that are needed to control the water distribution network, for example pumps increase the pressure load so that the water flows into the network. The Figure 1 presents a simple network with 1 resource “rs”, 1 tank “rv”, and 4 ordinary nodes “od1-4” of which “od4” is a consumer. An equipment pumps water from the resource into the network. The tank accumulates the water not consumed, or distributes water into the network to satisfy the demands. The aims of the different schemes that are presented hereafter are to model the hydraulics and the transport of chemical agents, but also to use those models for sensor placement and source identification.

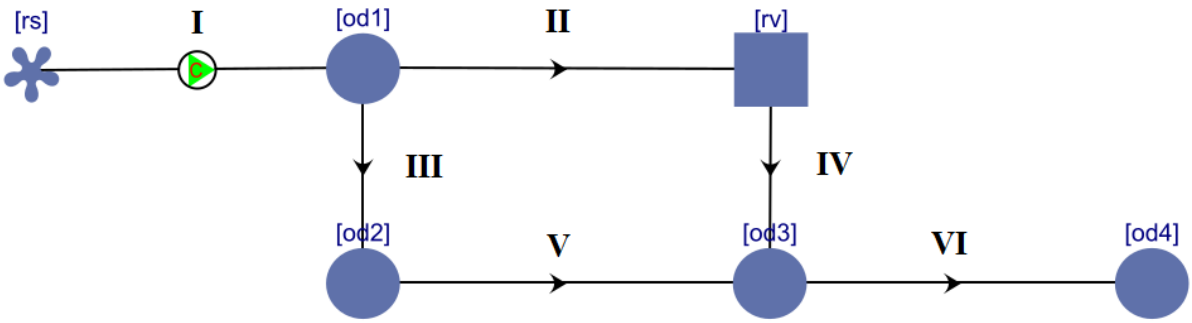


Figure 1: Water distribution network example.

I-1 Hydraulic Model

The hydraulics in a water distribution network represents the dynamic state of the fluid in the network. It is defined by different variables such as the velocity/flow rate in the pipes and the head at nodes. This state is generally computed as permanent by constant time step in the hydraulic model. The state at a new time step is computed by minimizing a function taking into account the energy calculated at the preceding time step.

I-1.1 Incidence Matrix

The incidence matrix represents the graph of the water distribution network. It is defined as follows:

$$A^N = \begin{cases} 1 \text{ if the pipe is going out from the node,} \\ -1 \text{ if the pipe is going into the node,} \\ 0 \text{ otherwise.} \end{cases}$$

Therefore, its size is node number × pipe number. It can also take a reduced form by simplifying the cycles of the graph as defined by Piller (1995). An example is given in Figure 2 giving the incidence matrix corresponding to the network of Figure 1.

	I	**II**	**III**	**IV**	**V**	**VI**
od1	-1	1	1	0	0	0
od2	0	0	-1	0	1	0
od3	0	0	0	-1	-1	1
od4	0	0	0	0	0	-1
rv	0	-1	0	1	0	0
rs	1	0	0	0	0	0

Figure 2: Incidence matrix.

Operations on its columns and rows permit to show characteristics of the graph such as the core, the trees and the bridges. It may be used to simplify the graph and therefore simplify the solving of the equations of the models, as described by Deuerlein et al. (2014). An example is given in Figure 4 where “rs”, “od1”, “od2” , “od3”, “rv” , ”I” , ”II”, “III”, “IV”, “V” form the core and the rest is a tree composed of one branch. “od2” and “od3” can also be merged, giving Figure 3. Also, trees such as the one composed by the branch “VI” and “od4” can be simplified on node “od3” by adding “od4” demands to “od3”.

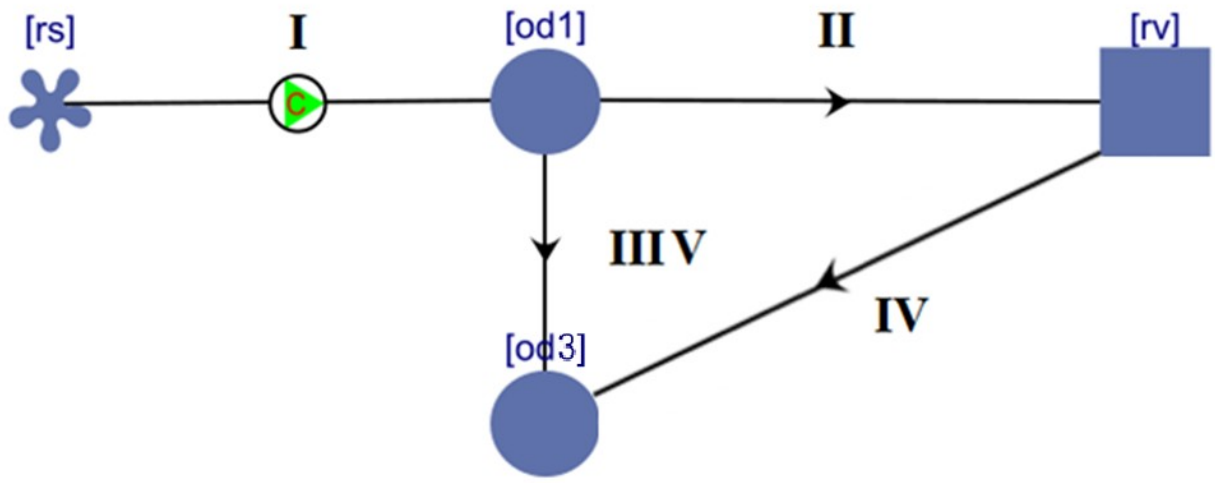


Figure 3: Water distribution network example simplified.

	I	II	IIIV	IV
rs	1	0	0	0
od1	-1	1	1	0
od3	0	0	-1	-1
rv	0	-1	0	1

Figure 4: Simplifications of the incidence matrix.

I-1.2 Analysis Problem and Content Model

Let us consider a pipe of length L and diameter D in which we consider a hydraulic scenario with the following characteristics: let q be the flowrates, d the demands, h the heads (composed of the pressure heads and the elevation heads), ξ the headloss function depending on q. The headloss functions are composed of friction headloss on pipes and singular headloss due to equipments, which are balanced with pump gains.

The mass conservation and energy conservation can be used to model the hydraulic behavior inside the network. According to the system given by the Analysis Problem (Piller, 1995), it is defined as follows:

$$\begin{cases} A.q = -d \\ \xi(q) - {}^{t}A.h = {}^{t}A^{f}.h^{f}, \end{cases} \quad (1)$$

with A the incidence matrix reduced to ordinary nodes and the notation ^{f} represents the fixed head nodes.

A typical headloss function for pipes is given by Hazen-Williams formula, (see Carlier (1980)):

$$\xi(r,q) = r|q|^{0.852}q, \quad (2)$$

where r, the resistance, takes into account the material of the pipe and is usually given by:

$$r = \frac{10.67\ L}{D^{4.87}CHW^{1.852}},$$

with L the length, D the diameter and CHW the Hazen Williams coefficient without dimension whose typical value is between 40 and 149 and is usually calibrated.

The Colebrook (1939) formula can also be used to find the Darcy friction factor λ.

$$\frac{1}{\sqrt{\lambda}} = -2\log\left[\frac{k}{3.7D} + \frac{2.51}{Re\sqrt{\lambda}}\right], \quad (3)$$

with k the roughness coefficient of the pipe and Re the Reynolds numbers. Numerical estimation is generally needed to find this parameter. λ is then used in the Darcy equation:

$$\xi\,(r,q) = \lambda\,\frac{L}{D}\frac{{u_0}^2}{2g}$$

with g is the local acceleration due to gravity and u_0 the average flow velocity.

The Analysis Problem can be solved with Piller (1995) method with the resolution of an optimization problem. The Content Model is defined as:

$$\min f(q) = \sum_{i \in pipes} \int_0^{q_i} \xi(q)\,dq - {}^{t}q\ {}^{t}A^{f}h^{f}.$$

This method consists in applying a Newton method to the equations with the (q, h) formulation:

$$F(q,h) = \begin{pmatrix} \xi(q) - {}^{t}A^{f}.h^{f} - {}^{t}A.h \\ A.q + d \end{pmatrix} = 0,$$

with Lagrange function:

$$L(q,h) = f(q) - {}^{t}h(Aq + d),$$

then: q^k and h^k known, q^{k+1} and h^{k+1} are computed through:

$$\begin{pmatrix} q^{k+1} \\ h^{k+1} \end{pmatrix} = \begin{pmatrix} q^k \\ h^k \end{pmatrix} - \begin{pmatrix} D_k & -{}^{t}A \\ A & 0_{nn} \end{pmatrix}^{-1} \begin{pmatrix} \xi(q^k) - {}^{t}A^f.h^f - {}^{t}A.h^k \\ A.q^k + d \end{pmatrix},$$

with D_k the Jacobian of ξ from q (a diagonal matrix).

Three Newton methods have been tested: Newton on the reduced problem, Newton on Lagrangian and Newton on dual problem. Multiple tests have shown that the Newton on Lagrangian has better performances than the other two and is now implemented in Porteau (Irstea hydraulic modeling software, 2016).

Let us remark that A is generally not a square matrix, the existence and uniqueness of a solution of the mass conservation is not guaranteed without consideration of the energy equation.

I-2 Transport Model

The solutions given by the hydraulic model are used in the transport equations applied to a chemical product injected inside the network. It is mostly used for chlorination injection propagation inside the network, which has been well motivated in France since the event of the year 2001. The transport modeling is a complex subject. Chemical agents can be found in two forms inside the network, either being freely transported and reacting in the bulk or participating in the formation of a biofilm. The biofilm is slowly built on top of the inside wall of the pipes. Depending on the hydraulic state, chemical product can be absorbed inside the biofilm or injected back in the network. Multiple works have been done in order to model that biofilm or to clean it off from the network. Sometimes, the incrustations can be so large that it is required to change the diameter of the pipe by a smaller one in the model. Here we will only get interested in the bulk transport and simple reaction with the biofilm.

I-2.1 Advection Reaction Equation and Perfect Mixing

The equation used for the transport of chemical products in a water distribution network is the advection reaction 1-D equation:

$$\frac{\partial C}{\partial t} + u\frac{\partial C}{\partial x} + KC^{\alpha} = 0. \qquad (4)$$

Additionally, the mixing condition at nodes is generally considered as perfect:

$$\forall \text{ node } i, C_{out(i)} = \frac{\sum q_{in(i)} C_{in(i)}}{\sum q_{in(i)}}. \quad (5)$$

Different models exist among the Eulerian methods and the Lagrangian methods. Two of them are presented: a hybrid Eulerian method and a Lagrangian method. Another one can be used for sensitivity analysis, the adjoint model.

The hybrid Eulerian model is the one used in the hydraulic free software Porteau of Irstea as described by Constans et al. (2003) and Fabrie et al. (2010). This is a time-driven scheme that works on a quality time step, which is smaller than the hydraulic time step. That quality time step dq is composed of two parts:

$$dq = dq_{MOC} + dq_{\theta}.$$

The calculation begins with a method of characteristics that solves exactly the transport on dq_{MOC}; and then use a $\theta -$ scheme (Crank-Nicolson if $\theta = 0.5$) on the other time step part dq_{θ}. The first part needs the utilization of an interpolation for the discretization of each pipe, therefore it may be source of errors and increases the calculation time.

The Lagrangian model is used in the open source software Epanet (2000) and has been developed by Boulos et al. (1994). This is an event-driven scheme which is determined by the boundary condition changes. The scheme transports segments of contaminant and manages the node mixing creating smaller mixed segments. A minimum size is necessary on those discretizations to avoid an explosion of the calculation time. The scheme uses the method of characteristics:

$$C(x,t) = C\left(x + dx, t + \frac{dx}{u}\right).$$

Wagner et al. (2015) have used the adjoint method and applied it to source identification. The adjoint method permits to focus on the sensors instead of the source of contamination. Those are mostly used to calculate sensibilities. The adjoint equations are given by

$$\begin{cases} \tau = T - t \\ X = L - x, \end{cases}$$

therefore

$$\begin{cases} \frac{\partial C}{\partial \tau} + u\frac{\partial C}{\partial X} - KC^{\alpha} = 0 \\ C(\tau = 0) = C(\tau = T), \end{cases} \quad (6)$$

and the previous calculation methods can be used again with few changes of signs.

An important need for operators is to have a precise quality model. Indeed, all simulations of security breakthrough come with the calculation of the propagation of a contamination. Errors on the quality model will affect the sensors responses simulated as well as the source identification and the evaluation of the propagation. Two enhancements to the current model are proposed. One is to add a model of imperfect mixing at double T-junctions, the other is to take into account the 3-D phenomena that are the velocity profile and the radial diffusion in a dispersion model.

For details on the implementation, see Appendix A.

I-2.2 Imperfect Mixing

There is a need, for water distribution network security, to have a reliable transport model in order to precisely predict the propagation of a contaminant. Two enhancements of the model are possible and one concerns the mixing at the junctions. Indeed, current model used to model the mixing at the junction with a perfect and instantaneous mixing condition. The outflow concentration is the flow rate weighted sum of the inlet concentration. However that assumption is not always actually verified.

In recent studies, Ho et al. (2008) and Choi et al. (2008), showed that mixing at crosses and double T-junctions is important with regards to security. It is therefore essential to add it when calculating the hydraulic quality state of the network. Ho et al. (2008) developed, and adjusted with experiments, the Bulk-Advective Mixing model BAM for cross-junction imperfect mixing depending on the Reynolds numbers at the inlets and outlets.

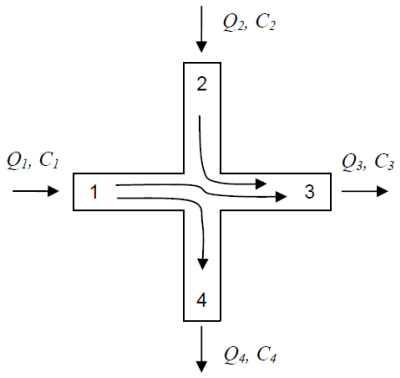


Figure 5: Cross configuration (from Ho et al. (2008)) with 2 inlets and 2 outlets.

In the cross-junction configuration of Figure 5, there is a principal flow 1, where the inlet goes into the two outlet directions 3 and 4 and the secondary flow where the flow inlet 2 is only going to the nearest outlet 3. For this configuration and within steady state condition, the pure BAM model reads:

$$\begin{cases} C_4 = C_1 \\ C_3 = \dfrac{Q_2 C_2 + (Q_1 - Q_4) C_1}{Q_3} \end{cases}, \qquad (7)$$

From Figure 5, it can be seen that $Q_4 < Q_1$, therefore $Q_1 - Q_4$ is positive. The complete BAM model consists of combining the perfect mixing and the pure BAM model with the following equation:

$$\mathrm{Cscaled} = \mathrm{Cbulk} + \mathrm{s}(\mathrm{Cperfect} - \mathrm{Cbulk}),$$

where s is a coefficient that depends on the real state of the cross-junction, Cperfect being the usual result without imperfect mixing and Cbulk is computed with Eq. (7).

They proposed to calibrate the s coefficient from experiments. Their results for X and N-junctions (or double T-junctions) are implemented in the Epanet-BAM module.

The AZRED model of Choi et al. (2008) proposes the use of a lookup table with interpolations or extrapolations for inputs not in the table. It was validated on experimental values with a large range of flows. One limitation is the need for extrapolation for Reynolds numbers approaching zero or infinity. In this thesis, it is proposed to complete the two previous models. The approach is to use a Computational Fluid Dynamics (CFD) based method. Main results have been published in Braun et al. (2014) and Ung et al. (2014).

The transport model equations that are usually used in water distribution hydraulic models are the 1-D advection-reaction equation coupled with a perfect mixing law. A 2-D solving has been proposed in Waeytens et al. (2015) to better assess the mixing at the junction by applying adjoint method and solving the Stokes equations in laminar flow ($\mathrm{Re} < 2{,}000$).

The solution proposed here is to complete Choi et al. (2008) works. CFD simulations have been launched in laminar and turbulent cases to create a lookup table upon which the high order interpolation Kriging is applied to create a 1-D imperfect mixing law at double T-junctions. A method based on the Delaunay triangulation is used in the design of experiment phase to get the next most relevant simulation to perform. The law created has been done with diameters all equal for the double T-junctions but has also been extrapolated for different diameters.

For details on the implementation, see Appendix B-I.

I-2.3 Dispersion

The other enhancement to the transport model concerns the propagation inside the pipe. The usual transport equation used is the advection reaction 1-D equation with the velocity taken as the average velocity inside the pipe. The dispersion modeling consists in both taking into account the velocity profile and the diffusion in the cross section of pipe.

Dispersion literature is quite recent, but most papers cite the article done by Taylor (1953) concerning the transport of a chemical agent in a pipe considering a velocity profile. He shows that a chemical agent that is placed inside a pipe is dragged by the fluid through the form of a diffusion propagation with a virtual diffusion coefficient E^*:

$$\frac{\partial C}{\partial t} + u_0 \frac{\partial C}{\partial x} - E^* \frac{\partial^2 C}{\partial x^2} + KC^{\alpha} = 0, \tag{7}$$

with C the concentration, u_0 the mean velocity, K and α the reaction coefficient and order, and

$$E^* = \frac{R^2 {u_0}^2}{48 D_m},$$

with R the radius of the pipe and D_m the molecular diffusion term in m^2/s.

This equation is however only justified for very long time:

$$T = \frac{D_m t}{R^2} \gg 0.5,$$

with T a dimensionless time.

Gill et al. (1970) extend the study also to small time by setting the virtual diffusion term dependent of the time E(t). They also proposed a method to calculate the exact solution of E(t) through the use of Bessel functions.

$$E(T) = D + E^*\left[1 - 768 \sum_{n=1}^{\infty} \frac{J_3(\lambda_n) J_2(\lambda_n)}{{\lambda_n}^5 J_0(\lambda_n)^2} \exp\left(-{\lambda_n}^2 T\right)\right], \tag{8}$$

with J_n the Bessel function of the first kind of order n and λ_n the root of J_1.

Lee (2004) has proposed in his thesis an approximation of it:

$$E(T) = D_m + E^* \left(1 - \exp\left(-\frac{T}{T_0}\right)\right), \quad (9)$$

with T_0 a Lagrangian term determined by calibration with Gill et al. (1970) expression, he has found the value

$$T_0 = \frac{1}{16},$$

then he deduces the average of that coefficient between 0 and t:

$$\overline{E(t)} = D_m + E^* \beta(T),$$

with

$$\beta(T) = \left(1 - \frac{1 - \exp\left(-\frac{T}{T_0}\right)}{\frac{T}{T_0}}\right).$$

That expression is approximated by Li (2006) for very small time ($T<0.01$):

$$\overline{E(t)} = \frac{u_0 L}{6}, \quad (10)$$

with L the length of the pipe.

However Romero-Gomez et al. (2009) show an overestimation of the coefficient given by Lee (2004) formula (Eq. (9)) and they propose a correction factor f_E which is around the value 0.7 in the cases they studied:

$$f_E = \frac{192(\frac{\overline{E^*}}{\nu})}{Re^2 Sc \beta(T)}, \quad (11)$$

with ν the cinematic viscosity, Re the Reynolds number and Sc the Schmidt number.

They have written in 2011 that for very small time ($T<0.01$) that coefficient is highly influenced by the boundary layer of the pipe. That pipe may not be axially homogenous and therefore they propose to use two coefficients β_f and β_b instead of one. Those would be calibrated with CFD simulations.

The method developed here models the dispersion effect as a 3-D phenomenon without virtual diffusion coefficient as defined in previous papers. 3-D CFD simulations being too time consuming, a random walk algorithm is applied coupled to a backtracking scheme to simulate the phenomenon. The backtracking scheme is an efficient way to get the results of particle transport

simulation. Indeed, all particles are launched at the location of the information, such as concentration, is wanted to be known. Therefore, fewer particles are needed to get the results. That scheme is well adapted to a coupling with a random walk scheme that models the radial diffusion.

For details on the implementation, see Appendix B-II.

I-3 Source Identification Inverse Problem

The enhancement of the transport model is necessary to precisely estimate the propagation of the contaminant. Some other needs of operators for the security of the water distribution networks are a source identification algorithm, which points out the locations and times of injection of the contamination source(s), and a sensor placement, which gives out the best locations for sensors to be placed in order to best monitor the network.

Literature shows different examples of contamination cases, accidental or intentional. One recent case happened at Laval (September 2015) where a malfunctioning wastewater treatment plant induced the dumping of river water directly in the water distribution network. One security measure had been to advise citizen to boil the water before consumption.

The source identification problem consists in finding the source of contamination, its starting time and the injection duration time for given contamination responses at sensors. This is an inverse problem and different methods exist to solve it. The source identification problem has been widely studied, and can be dispatched in three categories.

The first one is the enumeration type that consists in determining a subgroup of nodes that may be sources of contamination. For instance, Shang et al. (2002) use a particle backtracking algorithm (PBA) and compute the system in reverse time to find the list of the potential sources of contamination. De Sanctis et al. (2010) use the same algorithm extended to binary sensor response. Propato et al. (2007) and Propato et al. (2010) simulate all possible scenarios of contamination to create the matrix that links any source to any sensor. That matrix can be used to find the list of potential nodes of contamination.

The second type is the exploration type, which does the ranking. It uses the previous list and gives probabilities of each node of that group to be the true source of contamination. Guan et al. (2006) formulate a Least-Squares optimization problem. Laird et al. (2006) prefer to use a mixed-integer formulation to solve the source identification problem with concentration measurements at sensors. Liu et al. (2011) solve the same problem with the use of a logistic regression model, the work has also been done with binary sensors but results are less good. Propato et al. (2007) and Propato et al. (2010) use the matrix of contamination they created and

apply a minimum relative entropy method (MRE), to compute probabilities and confidence bounds. Also, Preis et al. (2006) couple Epanet (2000) with a genetic algorithm method.

Finally, the last type regroups the methods which take into account the sensor's error rate and use stochastic methods to calculate contamination source probabilities through the network. For instance, Dawsey et al. (2006) propose a Bayesian Belief Network (BNN). Also Perelman et al. (2010) developed a stochastic method on directed acyclic graph (DAG) which creates clusters of nodes upon whom the probabilities are then calculated.

The method developed here does both enumeration and exploration. It uses a backtracking algorithm to find the potential sources of contamination while creating the input/output matrix of contamination. The created matrix is the same as the one created by Propato et al. (2010) but not restricted to only positive responses, it is also extended to large networks. Indeed, Propato et al. (2010) method is direct and therefore needs to calculate all possible scenarios of contamination that may launch an alarm, which is time consuming. The reverse method proposed here uses sensor responses, either positive or negative, to tackle the adjoint equations. It takes considerably less time for large networks, which is important for a real time solving. That matrix will then be used to give rankings through counting positive values on rows and columns. The more a node can explain positive sensor responses, the more likely it is to be source of contamination. By crossing information on two nodes or more, it is also possible to define a probability for multi-contamination for each group of nodes it may come from.

The adjoint equation (see Eq (8)), presented before, allows the resolution of the source identification problem. Indeed, taking the sensor responses as initial conditions, the results of the adjoint equations give the possible solutions that may be sources of responses. This is a fast method to get the list of potential sources of contamination, however it does not produce the ranking for being the true source of contamination.

The backtracking model, used in this thesis, is similar to the previous method but is discrete. The previous one solves the transport equations with continuous methods, propagating the solution inside pipes. This one only cares about what is happening at the nodes, calculating the time to go from one end of a pipe to the other end. Particles are then sent back to get discrete information on the network state. This discrete form allows the creation of the input/output contamination matrix and gives a ranking through a counting algorithm. It is more detailed in chapter IV and it is also used for the sensor placement.

For details on the implementation, see Appendix C-I.

Also, the paper written by Seth et al. (2016) introduces an evaluation method to test the capabilities of a source identification method. They define two criteria, the accuracy and the specificity as follows:

$$\text{accuracy}\ (\%) = \frac{\text{Likeliness measure of the true injection node}}{\text{Highest likeliness measure over all candidate nodes}} \times 100,$$

$$\text{specificity}\ (\%) = \frac{\text{Number of nodes with lower likeliness than true injection node}}{\text{Total number of candidate nodes}} \times 100.$$

Two close definitions of those criteria are used in this thesis. The accuracy taken in this thesis is a binary value based on whether or not the true source of contamination is in the list of potential contamination nodes. It is the characteristic function of the value given by Seth et al. (2016) formula. The specificity is then defined in this thesis to only concern scenarios where the source identification is accurate. This criterion also includes the number of nodes with lower likeliness than the true injection node. However it is here divided by the number of all the nodes of the network instead of the number of potential contamination nodes. Those two new definitions are adapted to an evaluation of a large number of contamination scenarios. The accuracy percentage indicates the number of scenarios where the source of contamination is included in the potential source node list. Therefore a large accuracy limits the number of scenarios where the true source of contamination is not found. And the specificity focuses on the capability of the source identification method to give a good rank to the true source of contamination among all the nodes of the network.

I-4 Sensor Placement Model

Another important issue for water network managers is the sensor placement design. Installations cost, material and data managing are factors that define how operators should buy and set their sensors. Optimization design algorithms are necessary to make the best suited choices for each network and objectives.

To my knowledge, Lee and Deininger (1992) were the first to formulate an optimal sensor placement problem for the location of water quality monitoring stations on water distribution systems. Then multiple authors proposed their formulation and methods which are listed in Rathi et al. (2014). At first it was with one objective such as time to detection or the coverage. And then multi-objective methods appeared. In 2006, in Cincinnati, was organized the Battle of the Water Sensor Networks (BWSN) by Ostfeld et al. (2008) where algorithms have been proposed to solve the sensor placement problem. Among them is the resolution of the p-median facility location problem by Berry et al. (2006) or Krause et al. (2006). It uses the fact that the greedy algorithm based on a nondecreasing submodular function gives nearly optimal function. Propato et al. (2006) formulate a close formulation within a mixed integer linear programming (MILP) formulation.

A few tackled the problem of sensor placement optimizing the source identification criteria. Preis et al. (2006) perform a clustering of the network and use a genetic algorithm. Propato et al. (2007) propose a method that minimizes the set of solutions given by the matrix of contamination, improving the source identification. Tryby et al. (2010) carry on that work and propose a method that improves the regularization of the matrix. They have reached the conclusion that detection likelihood and source identification are correlated. Liu et al. (2014) use a multi-optimization genetic algorithm and draw Pareto fronts for these two criteria.

Determinist and stochastic methods look for the sensor locations (or placements) that minimize or maximize one or multiple criteria. The multi-criteria method developed by Irstea can be cited, see Piller et al. (2015). It uses the greedy method associated with a Monte Carlo method on different scenarios of contamination and criteria such as average time to detection or detection likelihood. A genetic algorithm can also be used, but most of the time it is highly time consuming. However it has the advantage to look for a global solution, avoiding local minima. The usual objective function criteria for sensor placement are: average time to detection, detection likelihood, contaminated population fraction, contaminated water volume. In this thesis a new criterion is defined on the rank of the potential node associated to the true source of contamination.

This thesis proposes new sensor placement methods that favor location helping the source identification process. The backtracking source identification method will be used to do the ranking. That ranking will serve as the sensor placement optimization objective.

A Monte Carlo method is associated with a greedy algorithm to calculate the best sensor locations. It looks, at each iteration, for the sensor location that maximizes the source identification criterion. Even though the method is quite costly in time, it can be parallelized. Also, another solution is used based on local search methods and graph theory which limit the sensor location search to the adjacent nodes of the current location estimations. From an initial guess of the best placement, the solution seeks placements that better comply with the source identification criterion. The search is done with three distance adjacent nodes through multiple iterations.

For details on the implementation, see Appendix C-II.

CHAPTER II

Enhancing the Transport Model: the Imperfect Mixing at Double T-junctions

For now, the transport of contaminants has been mostly modeled assuming perfect mixing conditions at T-junction. However, some studies have shown that it is not always the case when crosses or double T-junctions are involved. In this thesis, the imperfect mixing at double T-junction model is used, taking into account 3-D behavior, in order to construct a 1-D law that can be applied to 1-D transport models for water distribution networks. A four parameter Kriging interpolation based on the ratio of Reynolds number at the inputs and at the outputs, the average Reynolds number, and the length of the intermediate pipe of the double T-junction is implemented into the 1-D model. To efficiently build the process, the Delaunay triangulation is used in combination with an estimation of the interpolation error to define the next region of space where a simulation needs to be done.

Precedent works have studied crosses with simulations and experimentations, and double T-junctions with experimentations. Our approach consists in performing 3-D CFD simulations to estimate both laminar and turbulent dynamics at double T-junctions. They are performed by solving Navier-Stokes equations with either advection equations or mixture models. The results give the behavior of the mixing at double T-junctions at design points. To get the solution for parameter values not in the table, a Kriging interpolation, suitable for cloud point distribution, Wim et al. (2004) , is performed. First, a calibration of its parameters is needed and is solved with a maximum likelihood function. Then the solution of the interpolation is defined as a weighted sum of the CFD simulation results with those weights found through the resolution of a linear system. An example is then given to explain its application. The interpolation is also used to define the interpolation error that is used in the Delaunay triangulation method to find new design points. These are the center of the triangles with the highest error. An interpretation of the interpolated result surfaces is then given concerning the behavior of the imperfect mixing depending on the four parameters: length of the interpipe, the average Reynolds number and the input and output Reynolds number ratios, for double T-junctions composed of pipes with equal diameters. And finally, the results are generalized by a 1-D law that is valid for mixing at double T-junctions in water distribution network models.

1-D advection-reaction equations are usually used to simulate the propagation of physico-chemical agents inside the pipes of a water distribution network.

$$\frac{\partial C}{\partial t} + u\frac{\partial C}{\partial x} + KC^{\alpha} = 0,$$

with C the concentration of the agent, t the time, x the position, u the average velocity of the pipe and K and α, the kinetic constant and order of the reaction respectively.

Perfect mixing at T-junction law is usually used

$$C_{out} = \frac{\sum_{i \in entries} Q_i C_i}{\sum_{i \in entries} Q_i},$$

with C the concentration and Q the flow.

II-1 Double T-junction Configuration

A double T-junction (Figure 6) is a special kind of junction composed of an interpipe connecting two T-junctions. In case of two inflows on one side and two outflows on the other side, the mixing may not be complete depending on the four parameters: Reynolds number fraction at the inputs, Reynolds number fraction at the outputs, the average Reynolds number, and the length of the interpipe, noted L.

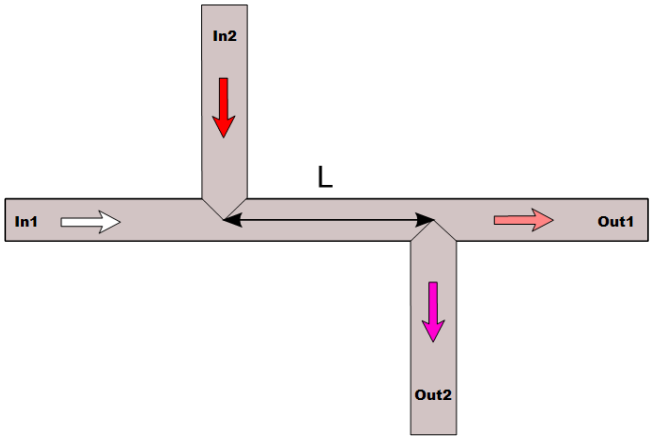


Figure 6: Double T-junction configuration with 2 inlets and 2 outlets.

It is proposed to use 3-D CFD simulations to create a 1-D law based on a lookup table and a 4-D Kriging interpolation method. Code Saturne, see Archambeau et al. (2015), has been used to simulate laminar flows and ANSYS Fluent for turbulent flows. Some common simulations have been completed for both software for turbulent cases and have given similar results, which justify that any of them can be used to model the phenomenon. They have been computed on the Mésocentre de Calcul Intensif Aquitain (MCIA) and Irstea computation grid respectively. The Jade supercomputer of Centre Informatique National de l'Enseignement Supérieur (CINES) has been used to assure the validity of the grid convergence.

The CFD simulations that have been carried out have been divided in 2 types:

The first one is composed of the laminar flow ($Re \leq 2000$) and transitional flow ($2000<Re < 4000$) cases. Direct Numerical Simulations (DNS) have been used and the transport model is the resolution of the advection equation for a concentration. In Figure 6, the boundary conditions are 0 mg/L for the straight inlet (In1) and 1 mg/L for the other one (In2). The velocities at 3 locations of the boundary are calculated from the laminar velocity profile formula:

$$u(r) = u_0\left(1 - \frac{r^2}{R^2}\right),$$

with u the velocity that is computed, u_0 the average velocity, r the radius from the center of the pipe and R the Radius of the pipe. The last boundary surface is pressure outflow set.

The second type is composed of the turbulent cases ($Re \geq 4000$). Large Eddy Simulation (LES) has been applied with Smagorinsky condition. The Mixture model of ANSYS Fluent has been used. Two fluids are defined to be water with similar properties but different initial conditions, the volume fraction of each fluid is set to 0 and 1 respectively at each input. Turbulent profiles have been developed and injected at the inlet and flow percentage is defined for each output.

II-2 Kriging Interpolation

An interpolation method is necessary to get the mixing with where simulations have not been performed. The simulations are not chosen from a Cartesian algorithm; therefore simple polynomial interpolations cannot be used. The interpolation needs to depend essentially on the distance from each simulation point. There exist several methods such as inverse distance interpolation or radial-based function interpolation. To make multi-dimensional inferences, we use in this research the well-known Kriging method or Gaussian Process Regression, see Chauvet (1982), Cressie (1993) and Jun et al. (2009) in geological sciences. It was created by Matheron and Krige. This is a minimum variance method based on the definition of a probability distribution at each of the n sample points x_i. It is based on the theory of regionalized variables and provides best results when data points are not evenly spread, which is our case. Compared to inverse distance interpolation, it gives more predominance to closer nodes as described by Wim et al. (2004). It also has the advantage of being more general than the radial-based function, which is a particular case treated by this method; see Costa et al. (1999). A comparison of 12 different types of interpolations is accomplished in Yang et al. (2004) and Kriging interpolation gives a very good approximation but is quite slow. However, here, the time of interpolation can be neglected in comparison with the CFD simulation time; therefore Kriging interpolation is a

good choice. The method chosen is the universal Kriging that we detail below. It consists in two steps.

II-2.1 Regression Model and Estimation of Parameters with Maximum Likelihood

In this section the aim is to resolve the calibrating optimization of Kriging parameters, considering n points of observation s_1, s_2,…, s_n. The function is taken as the sum of a polynomial function of s and a stochastic scalar function $Y(s)$.

$$Z(s) = Y(s) + \sum_{l=1}^{p} f_l(s)\beta_l, \tag{12}$$

with β_l the unknown coefficient of the polynomial.

We set:

$$F(s) = (f_1(s),..,f_p(s)) \text{ and } \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix},$$

therefore,

$$Z(s) = Y(s) + F(s)\beta.$$

Remark: if the row-vector $F(s)$ is only composed of one element equal to 1, the method is called ordinary Kriging, otherwise it is universal Kriging.

It is also supposed that the stochastic part is of expectation zero.

$$E[Y(s)] = 0.$$

It implies:

$$Cov(Z) = Cov(Y) = E[YY^T],$$

with

$$E[Z(s)] = F(s)\beta,$$

where Z and Y are the column n-vectors with components $(Z(s_i))$ and $(Y(s_i))$.

Finally the covariance is set as:

$$K = \mathrm{Cov}(Y) = \sigma^2 R(Y),$$
$$R\left(Y(s_i), Y\left(s_j\right)\right) = \prod_{k=1}^{m} \exp\left(-\theta_k \left|s_{ik} - s_{jk}\right|^{\gamma_k}\right), \tag{13}$$

where m is the dimension of the problem (here 4) and θ_k and γ_k are to be estimated.

The same equation is used when calculating $R(Y(s_0),Y(sj))$ with the coordinates of the estimation point s_0 instead of any observation point s_i.

It is supposed that the probability density function associated to Y on the sample points s_1, s_2,…, s_n is a multidimensional Gaussian:

$$\frac{1}{(2\pi)^{\frac{n}{2}} \det(K)^{\frac{1}{2}}} e^{-\frac{1}{2}(Z-F\beta)^T K^{-1}(Z-F\beta)},$$

with

$$K = (K_{ij})_{i,j\in[1,n]}, Z = (Z(s_i))_{i\in[1,n]}, \text{et } F = (F_l(s_i))_{i\in[1,n],l\in[1,p]}.$$

The Y probability density function can be rewritten as:

$$\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}} \det(R)^{\frac{1}{2}}} e^{-\frac{1}{2\sigma^2}(Z-F\beta)^T R^{-1}(Z-F\beta)},$$

with

$$R = (R\left(Y(s_i), Y\left(s_j\right)\right))_{i,j\in[1,n]},$$

and $\sigma, \beta, \theta_k\, \gamma_k$ $(k \in [1, m])$ are estimated with the help of a maximum likelihood optimization.

The function log-likelihood to minimize is:

$$f(\sigma, \beta, \theta_k, \gamma_k) = \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2}\log(\det(R)) + \frac{1}{2\sigma^2}(Z - F\beta)^T R^{-1}(Z - F\beta).$$

The function f is differentiated with respect to σ and β to derive necessary optimality condition by cancelling the gradient function:

$$\frac{\partial f(\sigma, \beta, \theta_k, \gamma_k)}{\partial \sigma} = \frac{n}{\sigma} - \frac{1}{\sigma^3}(Z - F\beta)^T R^{-1}(Z - F\beta),$$

therefore,

$$\frac{\partial f(\sigma, \beta, \theta_k, \gamma_k)}{\partial \sigma} = 0 \Rightarrow \hat{\sigma}^2 = \frac{1}{n}(Z - F\beta)^T R^{-1}(Z - F\beta) > 0, \tag{14}$$

and

$$\nabla_{\beta} f(\sigma, \beta, \theta_k, \gamma_k) = \frac{1}{2\sigma^2} 2F^T R^{-1}(Z - F\beta),$$

then

$$\frac{\partial f(\sigma, \beta, \theta_k, \gamma_k)}{\partial \beta} = 0 \Rightarrow \hat{\beta} = (F^T R^{-1} F)^{-1} F^T R^{-1} Z. \quad (15)$$

The gradients of function f with respect to θ_k and γ_k are given by:

$$\frac{\partial f(\sigma, \beta, \theta_k, \gamma_k)}{\partial \theta_k} = \frac{1}{2} \mathrm{tr}(R^{-1}[R \circ D_k]) - \frac{1}{2\sigma^2} (Z - F\beta)^T R^{-1} [R \circ D_k](Z - F\beta),$$

where tr is the matrix operator trace;

$$[R \circ D_k]_{ij} = R_{ij} \times (D_k)_{ij},$$

with o the product of Hadamard; and

$$(D_k)_{ij} = -|s_{ik} - s_{jk}|^{\gamma_k},$$

$$\frac{\partial f(\sigma, \beta, \theta_k, \gamma_k)}{\partial \gamma_k} = \frac{1}{2} \mathrm{tr}(R^{-1}[R \circ E_k]) - \frac{1}{2\sigma^2} (Z - F\beta)^T R^{-1} [R \circ E_k](Z - F\beta),$$

where

$$(E_k)_{ij} = -\theta_k \log(|s_{ik} - s_{jk}|) |s_{ik} - s_{jk}|^{\gamma_k},$$

Finally σ, β, θ_k γ_k are estimated by solving the following minimization problem:

$$\begin{cases} min\, f(\sigma, \beta, \theta_k, \gamma_k) = \frac{n}{2} log(2\pi\sigma^2) + \frac{1}{2} log(det(R(\theta, \gamma))) + \frac{1}{2\sigma^2} (Z - F\beta)^T R(\theta, \gamma)^{-1} (Z - F\beta) \\ \sigma \in \mathbb{R}^+, \beta \in \mathbb{R}^{+p}, \theta \in \mathbb{R}^{+m}, \gamma \in \mathbb{R}^{+m}. \end{cases} \quad (16)$$

It is solved using the Trust Region Optimization as defined by Conn et al. (1988) by giving it the gradients which allows finding the minimum of an unconstrained multivariable function. From a first trust region it then expands or contracts the search region by comparing the predicted and actual improvement of the objective function realization. As the problem is not strictly convex and not defined everywhere (see Figure 7), simplifications of the problem are done. σ and β are taken as optimal, values of γ_k components have all been fixed to 1 and the

θ_k have been expressed with regards to a non-dimensional parameter scaled by the maximum distance in all four directions:

$$\theta_k = \frac{\tilde{\theta}}{\max\left(\left|s_{ik} - s_{jk}\right|\right)^{\gamma_k}},$$

with $\tilde{\theta}$ the non-dimensional parameter that needs to be optimized. A second one is the regularization term defined thereafter. In this case, and taking into account the observations collected, the objective function may be undefined (see Figure 17), due to the correlation matrix R for small θ, which is not an invertible matrix. Therefore a regularization term is added to the diagonal of the correlation matrix R:

$$R_\alpha = R + \alpha\, I_n,$$

with α the regularization term. This is equivalent to making a ridge regression or using a Tikhonov regularization technique, it is also called nugget effect.

The new problem to solve becomes:

$$\begin{cases} \min f(\tilde{\theta}) = \frac{n}{2}\log(2\pi\hat{\sigma}^2) + \frac{1}{2}\log\left(\det\left(R_\alpha(\theta)\right)\right) + \frac{1}{2\hat{\sigma}^2}\left(Z - F\hat{\beta}\right)^T R_\alpha(\theta)^{-1}\left(Z - F\hat{\beta}\right) \\ \tilde{\theta} > 0, \theta = \frac{\tilde{\theta}}{\max\left(\left|s_{ik} - s_{jk}\right|\right)^{\gamma_k}}. \end{cases} \quad (17)$$

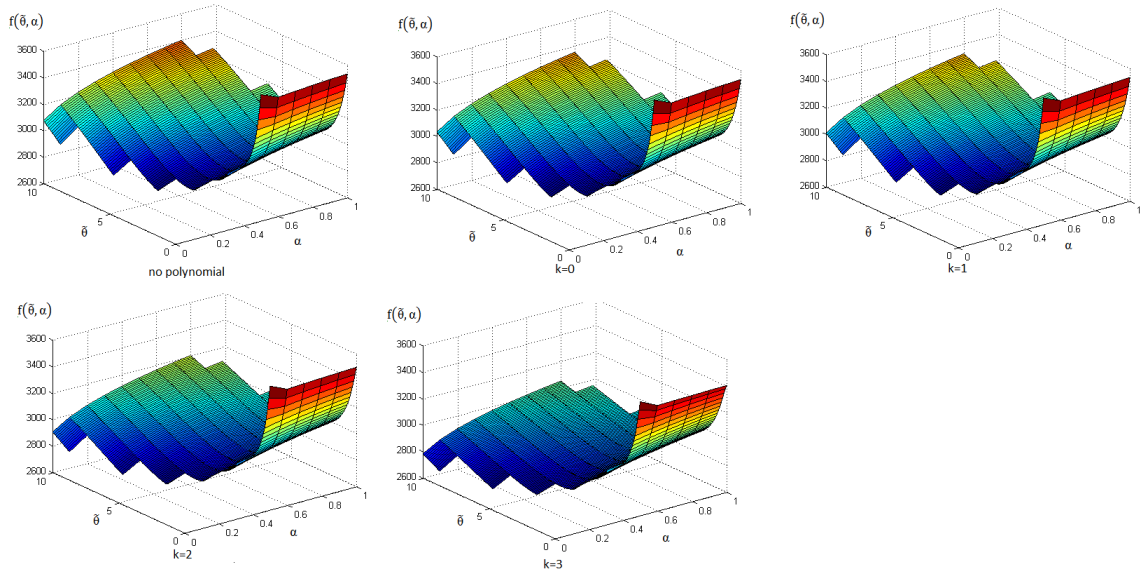


Figure 7: Contour plot of the log-likelihood function with respect to the shape parameter theta and the regularization parameter for polynomial equals to zero or its degree equals to 0, 1, 2 and 3.

The problem is not convex, therefore to understand the behavior of the function $f(\tilde{\theta}, \alpha)$, it has been ploted in Figure 7 with the generalized correlation parameter $\tilde{\theta} \in [0,10]$, the regularization term $\alpha \in [0,1]$ and the polynomial degree k between 0 and 3 as well as no polynomial, k being the degree of the polynomial. With σ, β fixed as optimal values from Eqs. (14) and (15). When there is not a polynomial term it is called simple Kriging (SK), when $k = 0$ its name is ordinary Kriging, for $k > 0$ it is universal Kriging.

For k between 0 and 3 and SK, there are 3 regions for α. If α is too small (<0.45), the R matrix may not be invertible (because it is not regularized enough), therefore its determinant is zero and the function f equals –infinity. On the other hand, if α is too large (>0.6), the correlation matrix may be too close to $(1+\alpha)I_n$, therefore its determinant is approximately $(1+\alpha)^n$, the matrix is invertible but its determinant in floating-point arithmetic is +infinity.

To get the optimized parameters, the system Eq. (17) has been resolved by choosing $\tilde{\theta} = 4$ at first iteration (because visually near the minimum) and $\alpha = 0.5$, for $k = 0, 1, 2, 3$:

Table 1: Kriging parameter optimization results by solving log-likelihood function for dimensionless $\tilde{\theta}$ and the regularization term α for a polynomial null or with a degree between 0 and 3.

Polynomial order	No polynomial	k=0	k=1	k=2	k=3
$\tilde{\theta}$	2.76	2.81	2.85	3.13	3.63
A	0.5	0.5	0.5	0.5	0.5
$f(\tilde{\theta}, \alpha)$ initial	2,902.92	2,895.64	2,891.3	2,855.29	2,806.2
$f(\tilde{\theta}, \alpha)$ optimal	2,866.93	2,864.77	2,863.32	2,841.27	2,804.62
Mean error	not exists	1.75	1.94	1.86	1.82
Standard deviation	not exists	3.40	3.24	2.83	2.37
Min	not exists	0.02	0	0	0
Max	not exists	35.19	34.09	31.4	26.86

The Table 1 sums up the result found by solving Eq. (17). Final value of $\tilde{\theta}$ increases with k. As σ, β are different for each k, the initial value of f gives different initial values. In this case the initial value of f is further away from the optimal value for lower values of k. The optimal value of f decreases with k. The average absolute error is the lowest for $k = 0$. Except for that, for average error and standard deviation, their values decrease with k increasing. Minimum errors are close to zero for all and maximum errors decrease with k increasing.

Finally, based on those results, we have chosen $k = 3, \hat{\sigma}, \hat{\beta}, \tilde{\theta} = 3.6333$ and $\alpha = 0.5$ with γ_k components all being fixed to 1.

II-2.2 Kriging Interpolation Solution

From the maximum-likelihood estimation of parameters σ, β, θ_k it is possible to predict the polynomial term of Eq. (12) or regression term. The residual term Y is not a white noise, it is possible to reduce it further by explaining the maximum of the intragroup variance.

An estimation of Z at the point s_0 is taken as a linear combination of Z from the observation points.

$$\hat{Z}(s_0) = \sum_i \lambda_i(s_0, s_i) Z(s_i), \tag{18}$$

where the λ_i are the unknowns of the problem.

Moreover we consider an unbiased predictor:

$$E\left[\hat{Z}(s_0)\right] = E[Z(s_0)],$$

that leads to:

$$\forall s_0, \sum_i \lambda_i F(s_i)\beta = F(s_0)\beta,$$

which is satisfied if:

$$\forall l \in [1, p], \sum_i \lambda_i f_l(s_i) = f_l(s_0).$$

We seek λ_i that minimizes the variance of the error term:

$$\min e(\lambda) = \mathrm{Var}\left[\hat{Z}(s_0) - Z(s_0)\right],$$

and by y expanding the error term:

$$\begin{aligned}\hat{Z}(s_0) - Z(s_0) &= \sum_i \lambda_i (Y(s_i) + F(s_i)\beta) - Y(s_0) - F(s_0)\beta \\ &= \sum_i \lambda_i Y(s_i) - Y(s_0),\end{aligned}$$

it follows:

$$\begin{aligned}Var\left[\hat{Z}(s_0) - Z(s_0)\right] &= Var\left[\sum_i \lambda_i Y(s_i) - Y(s_0)\right] \\ &= E\left[\left(\sum_i \lambda_i Y(s_i) - Y(s_0)\right)^2\right] - \left(E\left[\sum_i \lambda_i Y(s_i) - Y(s_0)\right]\right)^2\end{aligned}$$

$$= E\left[\left(\sum_i \lambda_i Y(s_i) - Y(s_0)\right)^2\right]$$

$$= E\left[\left(\sum_i \lambda_i Y(s_i)\right)^2\right] - 2\sum_i \lambda_i E[Y(s_i)Y(s_0)] + E\left[\left(Y(s_0)\right)^2\right]$$

$$= \sum_i \sum_j \lambda_i \lambda_i Cov\left(Y(s_i), Y(s_j)\right) - 2\sum_i \lambda_i Cov\left(Y(s_i), Y(s_0)\right) + Var\left(Y(s_0), Y(s_0)\right).$$

We recall that:

$$K_{ij} = \mathrm{Cov}\left(Y(s_i), Y(s_j)\right).$$

We have found that the function to minimize is quadratic in lambda with Hessian K that is symmetrical definite positive (under mild conditions of θ). It is therefore strongly convex.

The convex minimization problem states:

$$\begin{cases} \min \mathrm{Var}[\hat{Z}(s_0) - Z(s_0)] \\ \forall l \in [1, p], \sum_i \lambda_i f_l(s_i) = f_l(s_0). \end{cases} \quad (19)$$

It is possible to define a problem without constraint by defining the Lagrangian function and the Lagrange multipliers μ_l such that:

$$L(\lambda_i, \mu_l) = \sum_i \sum_j \lambda_i \lambda_i K_{ij} - 2\sum_i \lambda_i k_i(s_0) + k_{00} + 2\sum_l \mu_l \left(\sum_i \lambda_i f_l(s_i) - f_l(s_0)\right) \quad (20)$$

or

$$L(\lambda_i, \mu_l) = \langle \lambda | K\lambda \rangle - 2\langle \lambda | k(s_0) \rangle + \sigma^2 + 2\left\langle \mu \middle| \sum_i \lambda_i f(s_i) - f(s_0) \right\rangle,$$

with $k_i(s_0) = \mathrm{Cov}\left(Y(s_0), Y(s_i)\right)$ and $k_{00} = \mathrm{Cov}\left(Y(s_0), Y(s_0)\right) = \sigma^2$

The sufficient and necessary optimality conditions are:

$$\begin{cases} \forall l \in [1,p], \sum_i \hat{\lambda}_i f_l(s_i) = f_l(s_0) \\ \forall i \in [1 \dots n], \sum_j \hat{\lambda}_j K_{ij} - k_i(s_0) + \sum_l \hat{\mu}_l f_l(s_i) = 0. \end{cases} \quad (21)$$

These last equations can be written as the saddle point equation:

$$\begin{pmatrix} K & F \\ F^T & 0 \end{pmatrix} \begin{pmatrix} \hat{\lambda} \\ \hat{\mu} \end{pmatrix} = \begin{pmatrix} K_0 \\ F(s_0)^T \end{pmatrix}.$$

Using the Schur complement of K in the full matrix, it leads to the expression of optimal λ and μ:

$$\hat{\mu} = [F^T K^{-1} F]^{-1} (F^T K^{-1} K_0 - F(s_0)^T),$$

$$\hat{\lambda} = K^{-1}(K_0 - F\hat{\mu}),$$

Moreover we also obtain an estimation of the error variance by using Eq. (21) in the Eq. (20):

$$\mathrm{Var}\left[\hat{Z}(s_0) - Z(s_0)\right] = \hat{\sigma}^2 - \hat{\lambda} K_0 - \hat{\lambda}^T F \hat{\mu}.$$

The Kriging interpolation consists of two parts: The first one consists in estimating the parameters σ, β, θ_k at maximum-likelihood sense; the γ_k are fixed to 1s; a regularization term α is added to the diagonal of the correlation matrix. Finally an optimization problem is solved to find the parameters. The second part consists in solving the interpolation as a weighted sum of the observation results. It gives an effective way of interpolating non-structured points.

The interpolation is called in the 1-D law based on the lookup table of the results. It is also used to determine the errors of interpolation used with a Delaunay triangulation to choose new points of simulation from an initial set.

II-2.3 Kriging Interpolation and Imperfect Mixing

To better illustrate how Kriging interpolation has been used, an example of results in a case of imperfect mixing is presented. All results are given later where they are interpreted. The imperfect mixing law is computed from Table 2 of the simulations results. The Kriging method is used to interpolate any value inside the Reynolds and distance range. Extrapolation out of range is not recommended and more researches need to be carried out there. The Table 2 gives a non-exhaustive list of the simulations performed. The first column is the distance between T-junctions (in diameters), the second is the average Reynolds number (between the T-junctions), the third

column is the Reynolds percentage of inlet 1 (straight inlet see Figure 6) from the previous one, the fourth is for the outlet 1 (%In2 and %Out2 can be found from 100 - %In1 and 100 - %Out1 respectively). The last two columns give the repartition of contaminant for outlet 1 and 2 averaged in time (Z1 and Z2). Indeed, LES and DNS simulations are transient and therefore results need to be averaged for a sufficient period of time (e.g., several dozen minutes) when converged (see Figure 8).

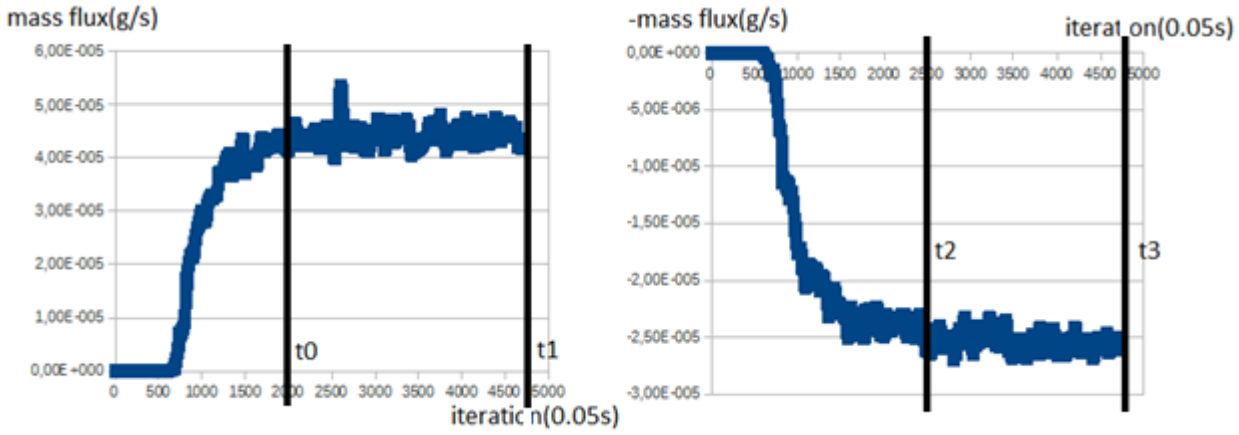


Figure 8: Example of simulation result: curve of mass flux (left: Out1, right: Out2) with averaging range.

Table 2: Results of the CFD simulations for different parameters (non-exhaustive).

Distance (in diameter)	Reynolds	%In1	%Out1	Z1	Z2
5	1,000	20	20	21.6	78.4
5	1,000	20	80	81.1	18.9
5	1,000	30	30	30.1	69.9
5	1,000	30	70	71.7	28.3
5	1,000	70	30	58.8	41.2
5	1,000	70	70	93.4	6.6
5	1,000	80	20	54.8	45.2
5	1,000	80	80	86.3	13.7
5	5,000	20	20	20	80
5	5,000	20	80	79	21
5	5,000	30	30	29	71
5	5,000	30	50	49	51
5	5,000	30	70	69	31
5	5,000	50	30	37	63
5	5,000	50	50	60	40
5	5,000	50	70	80	20
5	5,000	70	30	44	56
5	5,000	70	50	66	34

Distance (in diameter)	Reynolds	%In1	%Out1	Z1	Z2
5	10,000	50	50	62.5	37.5
5	10,000	50	70	82	18
5	10,000	70	50	69	31
8	5,000	30	30	31	69
8	5,000	30	50	50	50
8	5,000	30	70	70	30
10	1,000	20	20	21.5	78.5
10	1,000	20	80	81.5	18.5
10	1,000	30	30	32.5	67.5
10	1,000	30	70	70.5	29.5
10	5,000	0	0	30	70
10	5,000	0	0	70	30
10	5,000	0	0	40	60
10	5,000	0	0	81.5	18.5
10	5,000	0	0	64	36
10	5,000	0	0	50	50
10	5,000	0	0	74	26
10	5,000	0	0	34	66

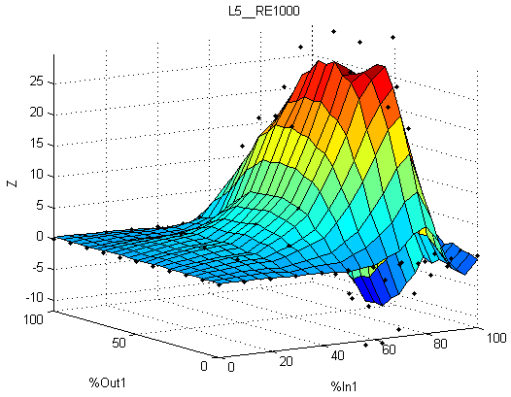


Figure 9: Interpolation result in case of distance = 5D and Reynolds number = 1,000.

For the interpretation of the results, it is worth recalling that 100% of the contaminant comes from inlet 2 and 0% from inlet 1. One way of interpreting the figures is by fixing two parameters and plotting the results with the other two varying.

In Figure 9, the distance is fixed to 5D and the Reynolds number is 1,000. It shows Z, the difference between the result found in the simulations and the complete mixing plan (Z=%Out1). Therefore the values are inside the boundary space [0 - %Out1; 100 - %Out1]. The difference is up to 30%, and therefore can't be neglected. It can also be noticed that for %In1<50, which is when the straight inlet is dominated, the mixing law is close to complete. The mixing is inversed when %In1>85 :more concentration is detected at Outlet 2 than in perfect mixing condition. A peak is visible with its top around %In1=70 found in every configuration.

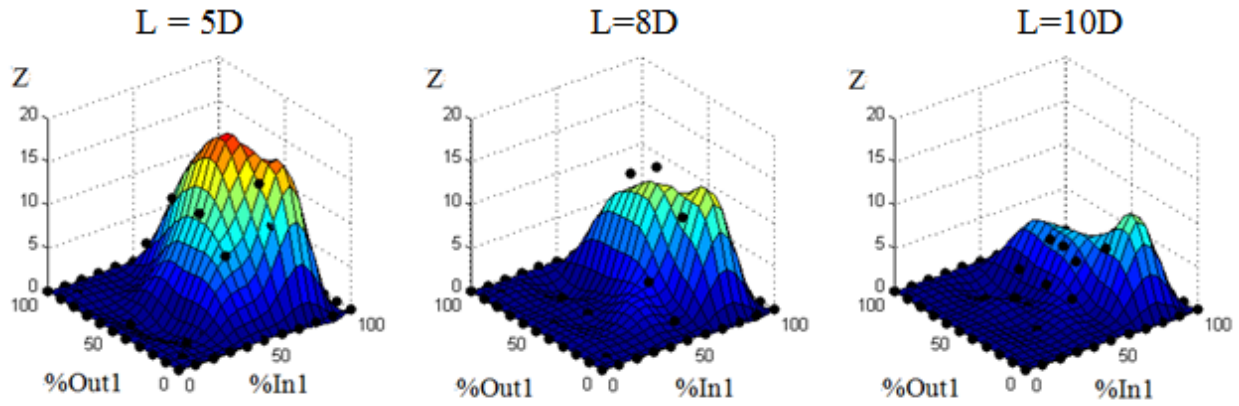


Figure 10: Comparison of interpolation results in case of Reynolds number = 5,000.

Figure 10 shows the results when the Reynolds number is fixed at 5,000 and the distance is increased from 5D to 10D. It can be observed that there is a peak situated in the right part (when the straight inlet is dominating). The effect of increasing the distance between the two T-junctions is a reduction of the peak in height and width.

An important issue is the choice of the boundary conditions: contrary to previous studies, the boundary conditions are fixed as the complete mixing case values. Indeed, a previous study Ho et al. (2008) also took these conditions for the left, bottom and top conditions but not for the right boundary condition ($\%In1 = 100$). Around this last one the inlet 2 is almost closed and therefore almost no contaminant enters, its repartition at the outlet is then not clear. The simulations tend to show that it is close to complete mixing state and therefore it was fixed in this study as such.

II-3 Design Plan and Delaunay Triangulation

One limitation of the study has been the limited number of simulations to be performed. Indeed, DNS and LES simulations are computationally demanding, therefore a method is needed to find the best points of simulation to perform. One well-known method is the Latin-Square design (see Abraham (1943)) consisting in dividing the space into a fixed number of squares and then randomly picking points in each square. However, for computational simulation, a major drawback is that it requires a lot of simulation points, at least one for each region defined. In this study we have selected the Delaunay triangulation method. This method has been used for the FFAST project by Bergmann et al. (2012) to improve a domain-decomposition. It was made to enrich the database of principal component analysis basis functions used in oscillating airfoils in a compressible flow context. Points of simulation could have been chosen from structured method or Latin Hypercube method (see Kay et al. (1979)) which is a generalization of the Latin-Square design, however more points of simulation would have been needed. The advantage of the method is to give a non-random optimized method for design of experiments.

It was determined in a previous study of Ho et al. (2008) that for $Re>10{,}000$, no further change in behavior is observed. Additionally, it has been shown by Ho et al. (2008) that after 20 diameters for the length between 2 T-junctions, complete mixing occurs. A discussion with operating partners of SMaRT-OnlineWDN has permitted to fix the low boundary for distance between T-junctions. Indeed, crosses and double T-junctions can't be compared easily, because for double T-junctions some space is needed to put a valve at the interpipe. That valve may also influence the computation but is, for now, not modelised. Therefore we have chosen the domain for the parameters described in the Table 3.

Table 3: Parameters domain.

Factors	Low level (-)	Mid level chosen	High level (+)
Reynolds average : Re	1,000	5,000	20,000
Distance inter-T : L	5×Diameters	10× Diameters	20× Diameters
Fraction Reynolds : %In1	0	50	100
Fraction Reynolds : %Out1	0	50	100

To complete the study, at least three experiments are needed for every factor. As CPU time is long to get the stability, 81 simulations should been carried out. However, we take advantage of simulation control points to get more reliable results in concordance with the Kriging interpolation and Delaunay triangulation.

The Delaunay triangulation method is usually used for space partitioning. It defines a list of triangles from a list of points where no points are strictly inside any circumcircle of any triangle. Here it is used to determine new points of simulation by a greedy algorithm:

- For each point of the design plan, we define the interpolation error as the absolute difference between its value and the interpolated value at that point when it is not taken into account;

- For every Delaunay triangle the sum of the interpolation error of its vertices multiplied by its area is computed;

- The best point candidates to include in the design plan are the centroid for triangles with the highest weight; then, in this research, the selected point is the closest point with rounded coordinates, which is more convenient for simulation.

A simple example is given in Figure 11 taking into account nine points, which are given in Table 4.The function M is the list of the measurements and E is the list of errors of interpolation when discarding the point.

Table 4: Example of the Delaunay method with associated errors.

M(x,y)	x = 0	x = 1	x = 2
y = 2	3	3	3
y = 1	2	2	2
y = 0	1	0	1

E(x,y)	x = 0	x = 1	x = 2
y = 2	0.70	0.48	0.70
y = 1	0.57	0.53	0.57
y = 0	1.75	1	1.75

In this case, the method divides the space in eight triangles of equal area. The maximum error is found at the points (0,0) and (2,0) equals to 1.75. In this case, the two-bottom triangles will most likely be chosen for new simulations (blue crosses), this is where more accuracy is needed. The method detects the most problematic point, here the only point not in the plan $z = y+1$, and refines around it.

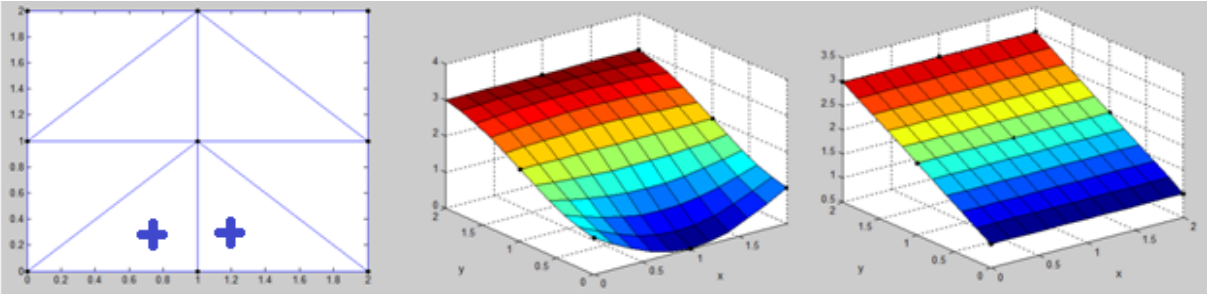


Figure 11: Result of the Delaunay method with: left) Delaunay triangles and new design points; middle) full interpolation; right) interpolation without point (1,0).

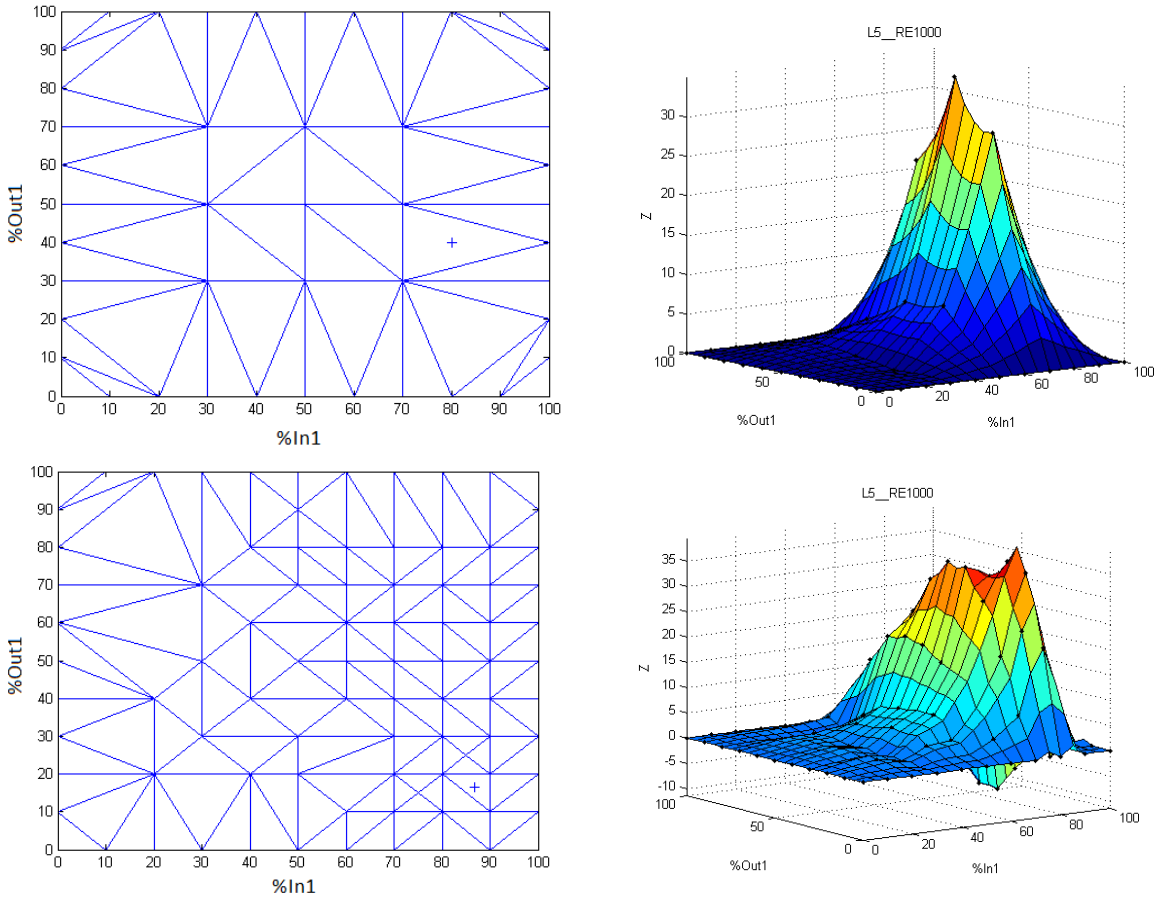


Figure 12: Example of Delaunay method application and interpolation of imperfect mixing for the distance = 5D, Reynolds = 1,000: up) before, down) after simulations chosen by the triangulation of Delaunay.

In Figure 12 the behavior of the Delaunay method is highlighted with an application of the case $L = 5D$ and $Re = 1{,}000$. On the left can be seen the Delaunay triangles partitioning the domain space, the vertices being the points of simulation or the boundaries. The blue cross defines the new point of simulation to perform. On the right is plotted the interpolation result, initially and after multiple simulations chosen by the Delaunay method. The repartition of the simulation chosen points is not structured, most points are chosen on the right part where the straight inlet is dominating, $\%In1 > 50$. It is also where the simulations give imperfect mixing results and the derivatives are bigger.

Table 5 sums up the different new simulation points to consider that were found for the case where $L = 5D$ and $Re = 1{,}000$ are fixed. The first column refers to the order of the simulations made given the Delaunay method and the number 0 is for the initial simulations. The second and third columns give the Reynolds inlet and outlet fractions compared to average Reynolds of 1,000. Finally the fourth and fifth columns are the sum of the errors, either averaged (divided by current number of points) or the maximum, of all the points simulated at each stage.

In Table 5 and Figure 13, it can be observed that for both types of error, the error is globally decreasing. The error can increase when a particular point is found, in this example (80, 30), but then the Delaunay method will search for points around that will decrease the global error. Figure 12 on the left shows the points of Delaunay locations (vertices of the Delaunay triangle). At the end the method concentrates the points where there is a need for more information, here in the right part, when the straight inlet is dominating.

Table 5: Delaunay points for distance = 5D and Re = 1,000.

Number	%In1	%Out1	Mean. error	Max error
0	30	30	5.1	14.1
0	30	70	5.1	14.1
0	50	50	5.1	14.1
0	70	30	5.1	14.1
0	70	70	5.1	14.1
0	50	30	5.1	14.1
0	50	70	5.1	14.1
0	30	50	5.1	14.1
0	70	50	5.1	14.1
1	80	40	5.2	14.7
2	80	60	4.9	13.9
3	60	60	5.1	14.3
4	60	40	5.3	11.9
5	80	50	4.5	11.9
6	60	80	4.4	11.1
7	80	30	3.9	13.3
8	80	20	4.1	12.4
9	70	20	4.1	12.4
10	80	80	4.2	12.4
11	70	80	3.7	12.4
12	70	10	3.6	13.3
13	90	10	3.8	9.8
14	80	10	3.8	9.8
15	40	60	3.7	9.6
16	70	60	3.3	12
17	90	20	3.3	12
18	50	20	3.3	12
19	50	90	3.0	8.2

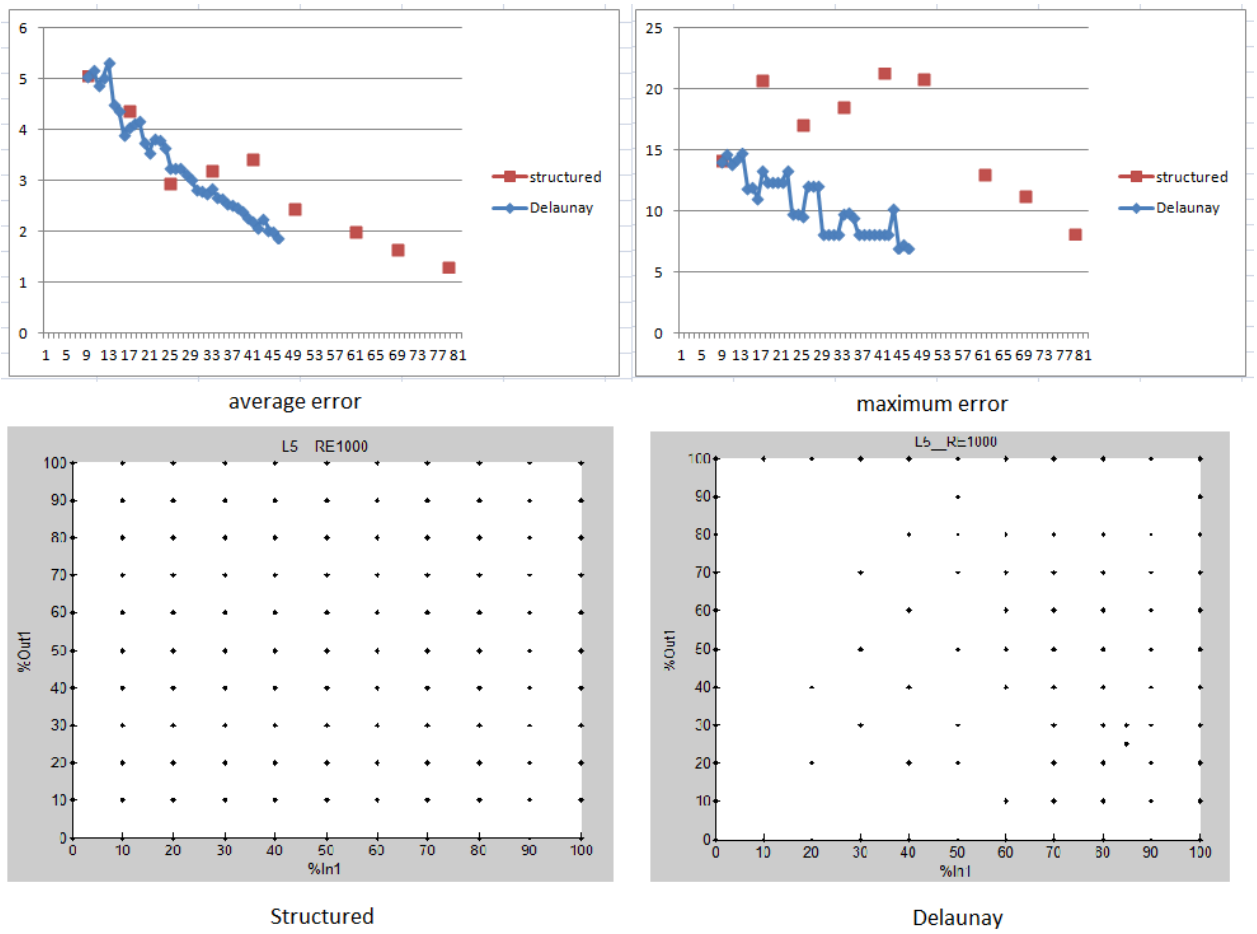


Figure 13: Comparison between structured and the Delaunay simulation point selection.

In order to show that the Delaunay method performs better than the Structured method in terms of design point selection, they are compared. Figure 13 shows the average and maximum error of interpolation (as explained before) when adding points of simulation. To compare both methods there is the need to define how the Structured algorithm was performed. The Structured selections have been made using the following sets of simulation points in Figure 14: the groups go from 0 to 8 (each having its own color). Each number represents a set of points, whose number is between 8 and 12 points. For instance, at first, both methods begin with the set 0 composed of 9 points corresponding to Reynolds number combination ratios of 30/50/70 Reynolds fraction at inlet and outlet. Then, 10/50/90 (set 1) combination is added to the precedent points. Henceforth it goes until set 8, when the domain is filled with a fully-structured repartition of points.

For both average and maximum interpolation error, the Delaunay method gives better results with a steady average behaviour of error decreasing. For the average error, both methods give similar results at first, until a point when the Delaunay method error stays smaller compared to the Structured error. For maximum error, the Structured method does not have a well-defined behaviour compared to average steady decrease of the Delaunay method. Each new increase peak

can be explained by new regions of space not well modelled. The Delaunay triangulation is then selecting points in that region until the error decreases again. This justifies the use of the Delaunay method over the Structured method.

%In1/%Out1	10	20	30	40	50	60	70	80	90
90	1	6	2	7	1	7	2	6	1
80	6	4	8	3	6	3	8	4	6
70	2	8	0	5	0	5	0	8	2
60	7	3	5	4	7	4	5	3	7
50	1	6	0	7	0	7	0	6	1
40	7	3	5	4	7	4	5	3	7
30	2	8	0	5	0	5	0	8	2
20	6	4	8	3	6	3	8	4	6
10	1	6	2	7	1	7	2	6	1

Figure 14: Structure selection points.

The Delaunay method has been used for space partitioning for selecting new CFD simulations to perform. At each step, it calculates the absolute interpolation errors, which can be used as a stop criterion. For instance, if we have fixed the mean error at 3 as the criteria of convergence, there is no need to continue the simulations after the 19th simulation. It gives an effective way to consume less computational time, by favoring simulations that give the most information.

II-4 Results

In this part CFD simulations are plotted and interpreted. First is given the repartition of simulations that have been carried out for the four parameters (Re, L, %In1 and %Out1). Then, two ways of interpreting the results are presented. One is by fixing the Reynolds number and the distance, the other by fixing the inlet and outlet Reynolds ratios.

The following notation is given XX/YY: XX% of the inlet flow comes from the inlet 1 (In1) and YY% of the outlet flow goes into the outlet 1 (Out1) of the double T-junction (see Figure 6).

All simulation points have been gathered in Figure 15 below:

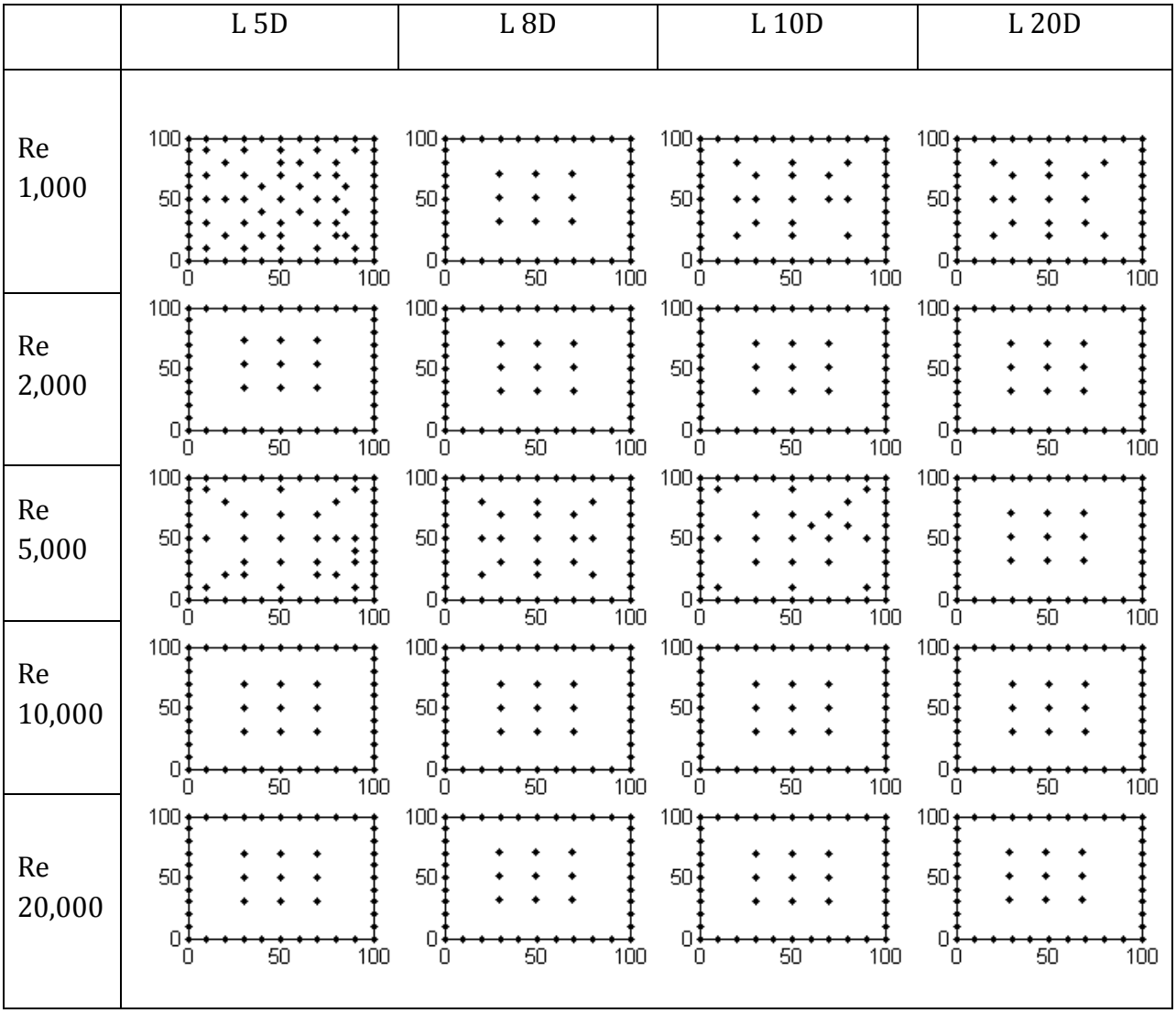


Figure 15: Points of simulations summary.

Figure 15 shows the list of simulations that have been realized, here represented by points. Each point is defined by its Reynolds number and its intermediate pipe length as well as its value of %In1 and %Out1 respectively in x-axis and y-axis. The Delaunay method has been mostly used on the space Re=1,000/L=5D and Re=5,000/L=5D to ensure the validity of the interpolation.

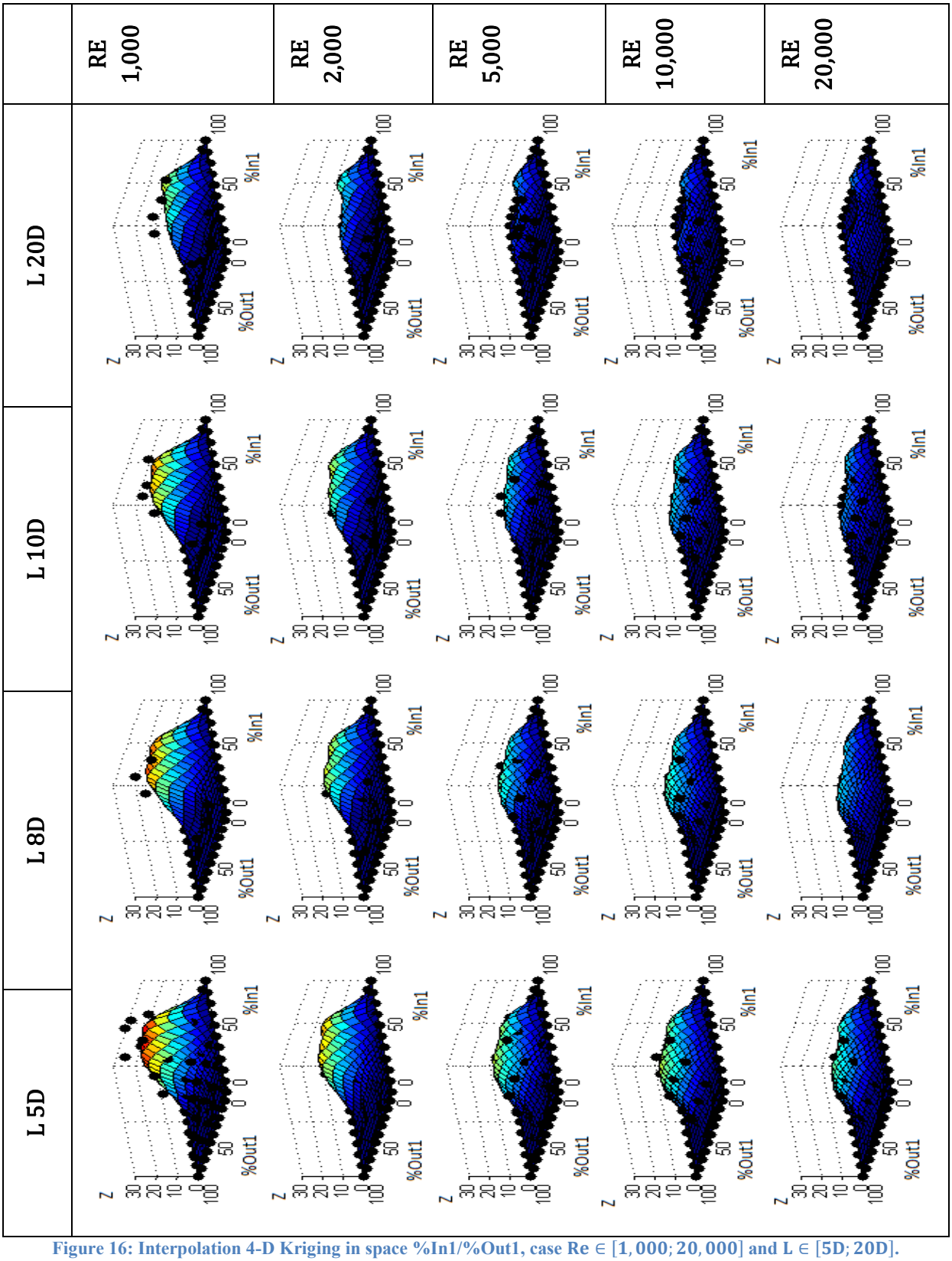


Figure 16: Interpolation 4-D Kriging in space %In1/%Out1, case Re ∈ [1,000; 20,000] and L ∈ [5D; 20D].

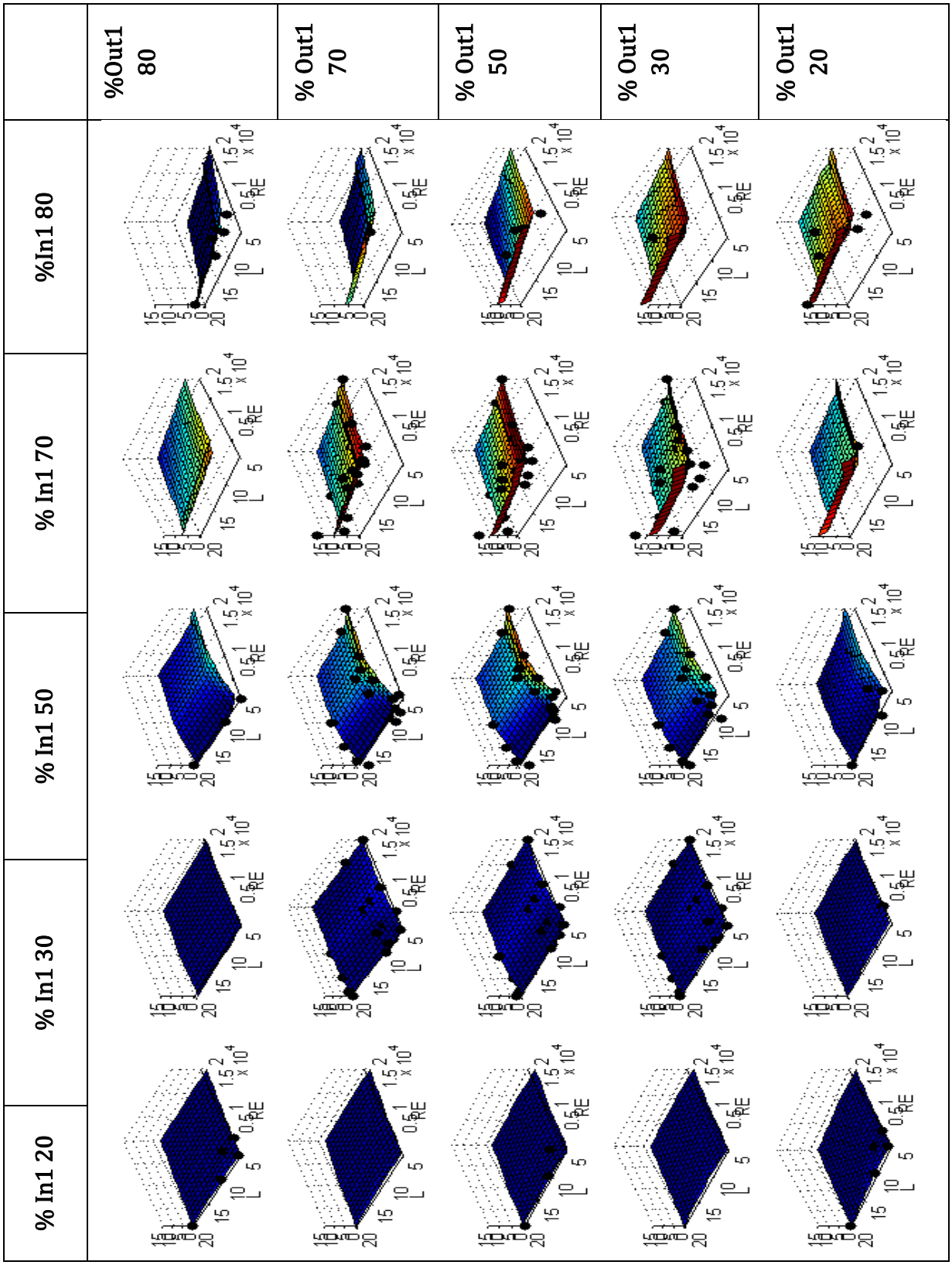


Figure 17: Interpolation 4-D Kriging in space L/Re, case %In1 and %Out1 = 20/30/50/70/80.

The results of the four parameter interpolation have been drawn in Figure 16 and Figure 17.

The first figure gives a lot of information on the behavior of this law. We may observe that when the second input (see Figure 6) is dominating ($\%In1 < 50\%$), the mixing is almost perfect. On the other hand, when it is the first input that is dominating, the mixing is not perfect. And it converges toward the perfect mixing with the increase of the distance and the Reynolds number. The case $Re = 30{,}000$, $L = 5D$, $\%In1 = 70$ was additionally computed and is showing that the function converges to a stable equilibrium for high Reynolds number, which may not be the perfect mixing depending on the distance. In case 5D, $70/50$ it tends to $Z = 10$, that means that in this case where the ratio at the output is 60% of mass introduced goes into Out1 (instead of 50%) and 40% in Out2. When D increases, a slow converging to perfect mixing in the laminar case can be seen. Figure 17 shows that for different configurations of Reynolds fractions in input and output, the behavior law is very different. As seen in Figure 16, when %In1 is lower than 50% the mixing is mostly perfect. In the laminar case the mixing has a clear behavior, under $\%In1 = 50$ it is perfect mixing and above 70 it is imperfect mixing. In the turbulent case it is more progressive.

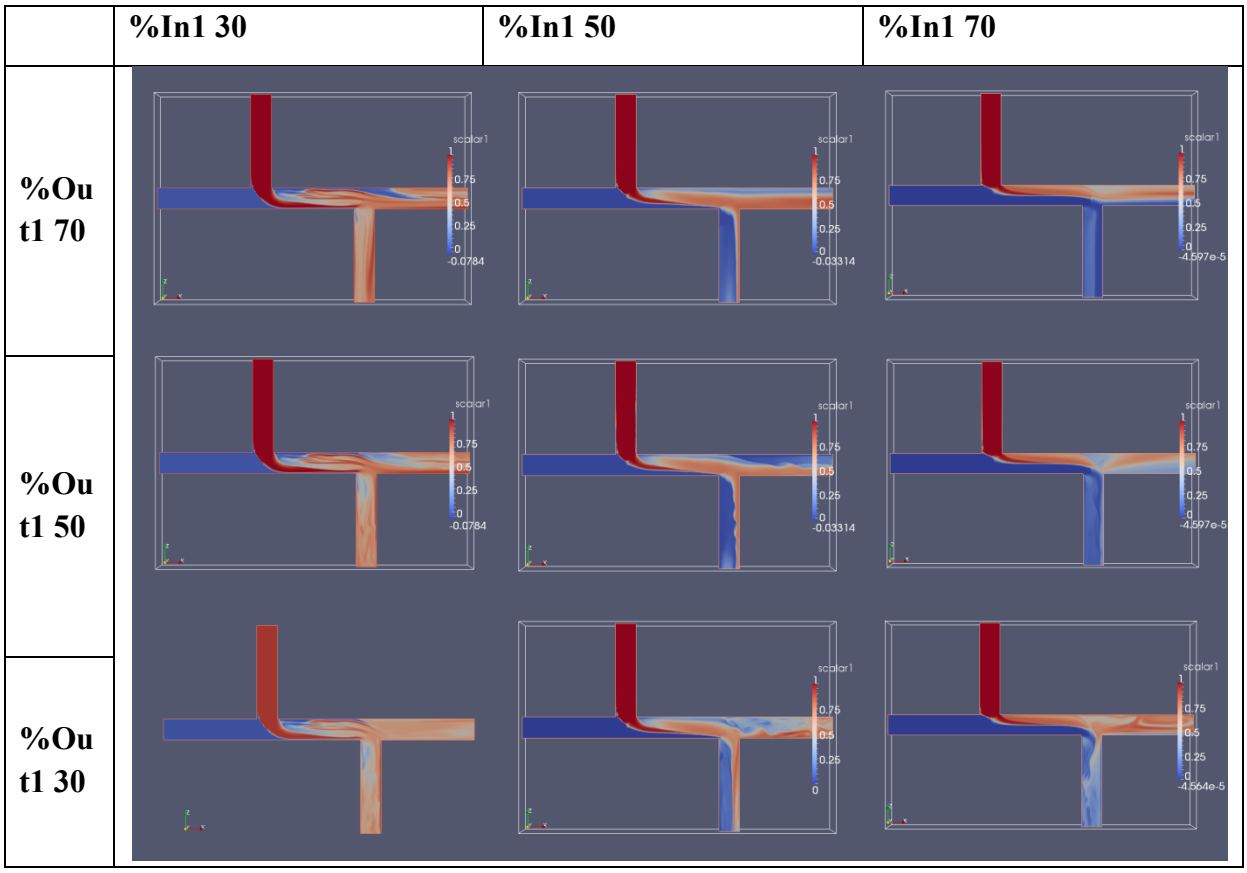


Figure 18: Representation of scalar on longitudinal section of Double T in cases Re = 1,000 fixed, L = 5D fixed, and %In1 and %Out1 = 30/50/70.

To understand the behaviour of the mixing, Figure 18 shows the longitudinal section of the double T-junction in the case $\mathrm{Re} = 1{,}000$, $\mathrm{L} = 5\mathrm{D}$ and $\%\mathrm{In1}$ and $\%\mathrm{Out1} = 30/50/70$ and plots the scalar. It can be seen that when the percentage of inlets at the first input is lower than $\%\mathrm{In1} = 30\%$, the contaminated water coming from In2 hits the opposite wall, creating turbulence. When $\%\mathrm{In1} = 50\%$, the turbulence is caused by the corner of the second T-junction. The more fluid going into the second output, the less stable is the flow. For $\%\mathrm{In1} = 70$, most of the contaminated water goes into the first output regardless of the output velocity repartition.

More than 250 simulations have been made to describe the phenomenon in four directions (Reynolds, inter T distance, $\%\mathrm{In1}$ and $\%\mathrm{Out1}$). The results have been interpolated and projected on two different spaces ($\%\mathrm{In1}$, $\%\mathrm{Out1}$) and (Re, L). When the first input is dominated ($\%\mathrm{In1} < 50$) the mixing is perfect. The mixing becomes imperfect when $\%\mathrm{In1}$ increases. This behaviour shift is more pronounced for laminar cases whatever the inter T-junction distance, and progressive in turbulent cases. There is a maximum of imperfect mixing around $\mathrm{In1}\% = 70$. The mixing tends to perfect mixing when the inter T-junction distance increases, it is almost perfect for $\mathrm{L} = 20\mathrm{D}$ in the turbulent case. It also decreases when the Reynolds number is increased but tends to equilibrium that may not be perfect mixing (tends to 10 when $\mathrm{L} = 5\mathrm{D}$, $\%\mathrm{In1} = 70$, $\%\mathrm{Out1} = 50$). The behaviour described can be explained by looking into the longitudinal section of the double T-junction. The mixing is created by hitting the walls, which depends on the ratios of the flows in input and output.

II-5 Experimentations

The test bench pilot is situated in Dresden at the TZW premises. It is shown in Figure 19.

Figure 19: Photo of the pilot in Dresden.

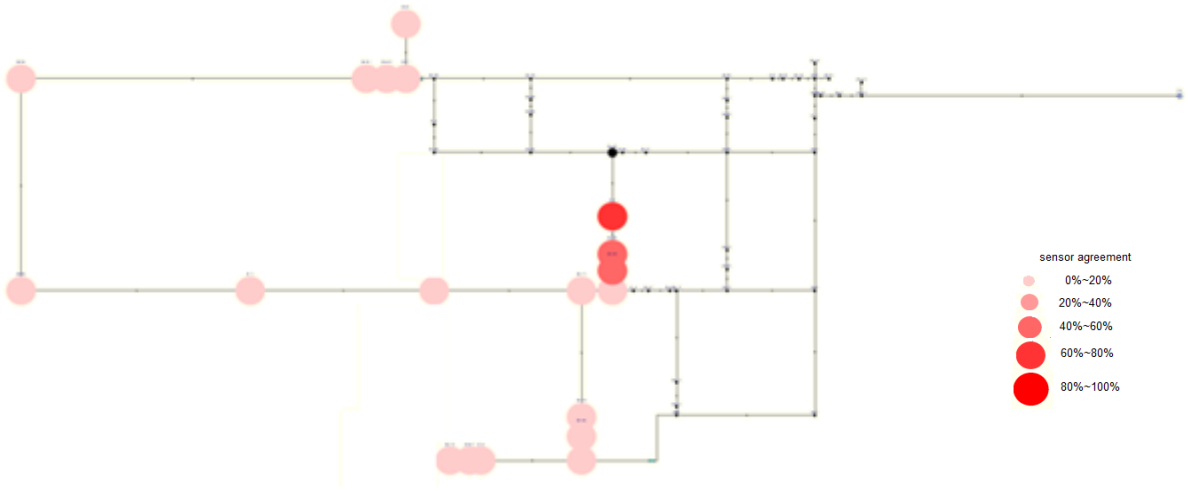


Figure 20: Schema of the pilot in Dresden with a contamination example.

Figure 20 shows an example of contamination in the network of TZW. A contamination is injected at the black point; then, the flow brings the contaminated water to a double T-junction downstream. The concentration of the fluid is then measured (on the left and bottom of the double T-junction) at the center of the tube by an electrical conductivity meter. Conductivity is used as surrogate for the concentration.

A comparison between the simulations and the experiment results has been made. Twelve experiment cases have been carried out: $L = 5D,\ 8D,\ 10D,\ 20D$. And for each case the following combinations of Reynolds 30/50/70 have been taken.

Figure 21 shows the comparisons between simulation and experiments for the laminar case, $\mathrm{Re} = 1{,}000$. The simulations and experiments show big differences in $70/XX$ cases when the flow going in inlet In1 is at maximum. It might be due to difficulties in keeping a stable low velocity because it is hard to measure. It might be also due to the property of the solution of contaminant. In the laminar case ($\mathrm{Re} = 1{,}000$) the contaminant travels more preferably to the bottom of the tube, the sensor being at the center of the tube.

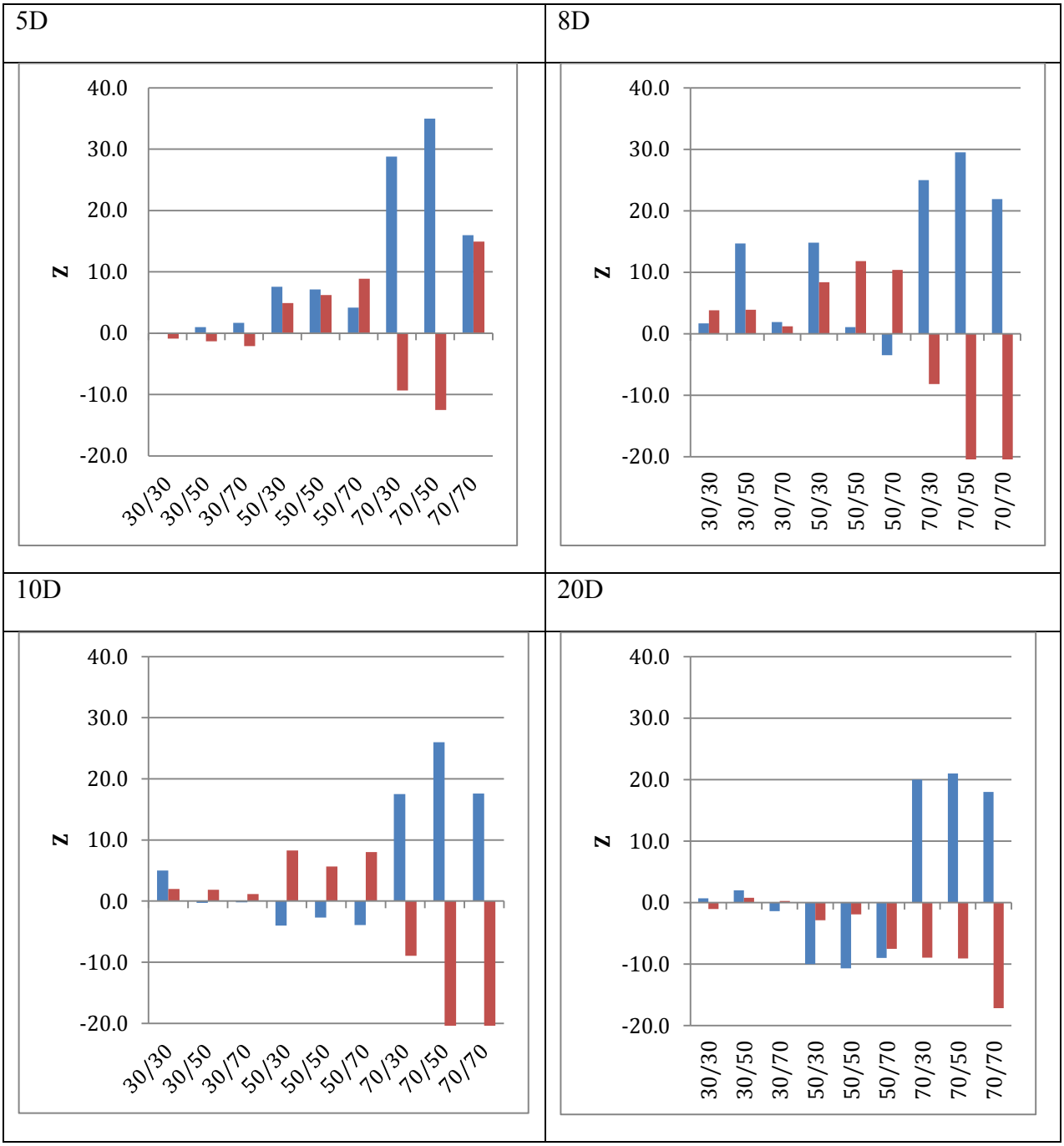


Figure 21: Comparison between CFD and experiments: RE = 1,000 (blue CFD simulations, red experiments).

Figure 22 shows better agreement, the trend of simulation and experiment is the same. It is almost perfect mixing when $\%In1 < 50$ and above zero for the cases 70/XX. Larger residues in the last cases may also be due to the contaminant falling slowly to the bottom.

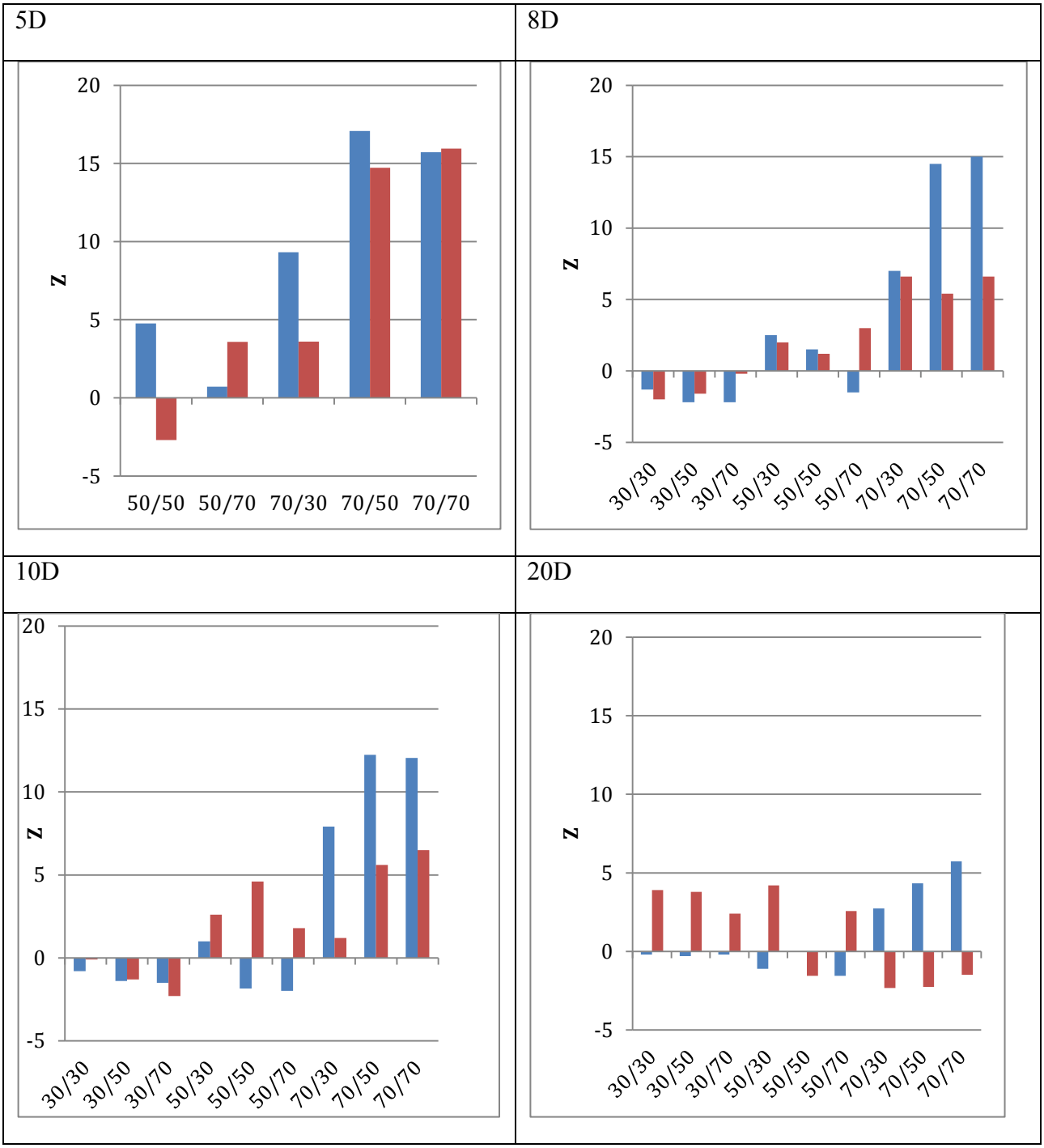


Figure 22: Comparison between CFD and experiments: RE = 2000 (blue CFD simulations, red experiments).

Figure 23 shows good matching between CFD simulations and experiments. The difference remaining might also be due to gravity. However as the fluid regime is turbulent, this can be caused by the hypotheses of simulation where the effect of turbulence is not taken into account enough.

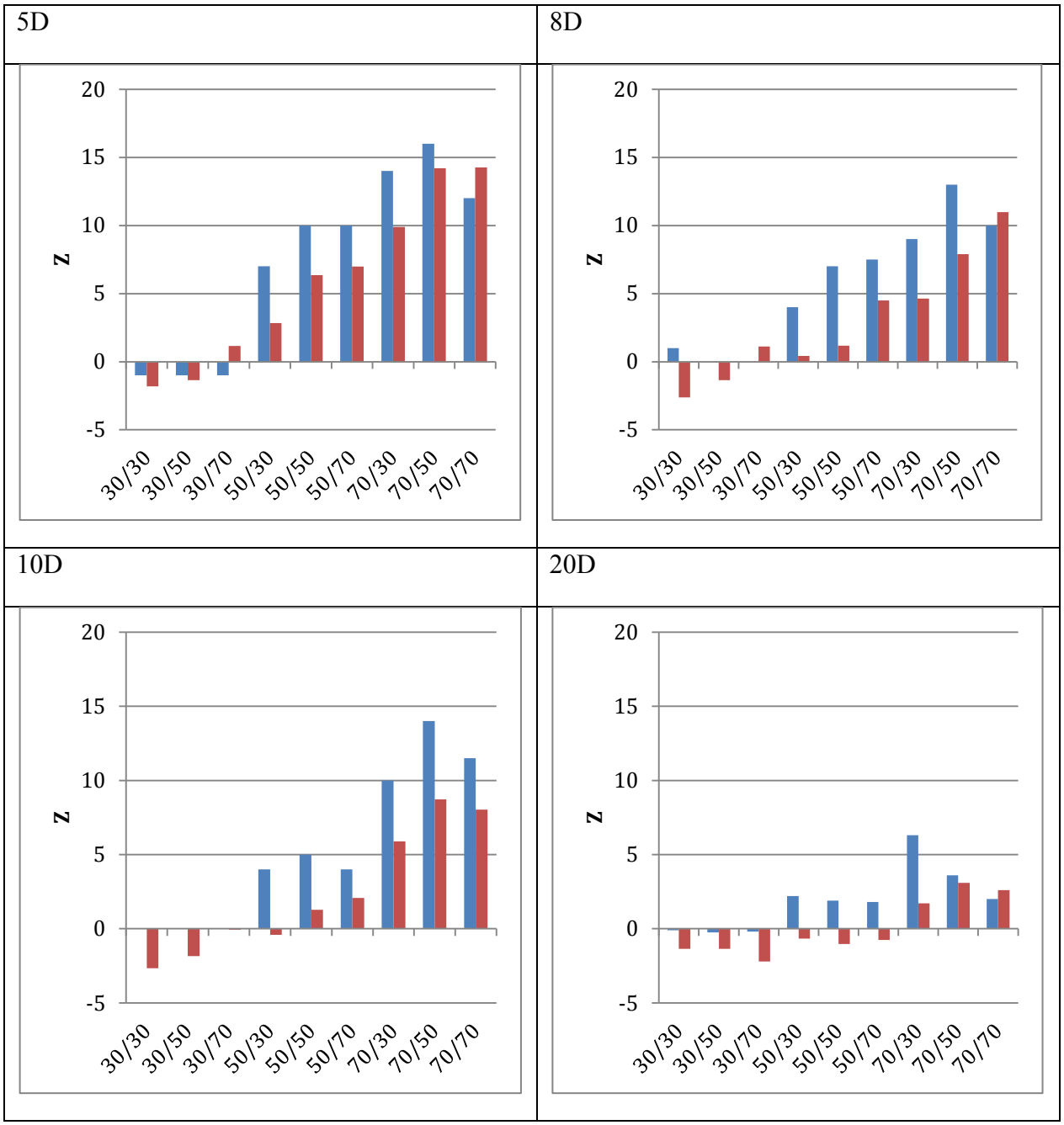


Figure 23: Comparison between CFD and experiments: RE = 5,000 (blue CFD simulations, red experiments).

In conclusion, the simulations and the experiments show a perfect mixing in the case where entry inlet 1 is dominating.

Some discrepancies between simulation and experimentation can be observed for the case $\mathrm{Re} = 1{,}000$, that may be due to the difficulty to measure very low flow in pipes.

A good matching can be seen in the cases $\mathrm{Re} = 2{,}000$ and $\mathrm{Re} = 5{,}000$. The simulations show however an overshooting compared to the experiments and may underestimate the turbulence.

II-6 1-D Law

A 1-D law can now be developed. It uses the lookup table of the simulations results and determines the proportion of mixing with the help of the Kriging interpolation. The simulations were executed based on the following scenario: pure water coming from In1 and contaminated water coming from In2. The law first needs to be generalized in case of any water concentration at each inlet.

Let us consider C_1, C_2, C_3 and C_4 the mean concentrations on the cross sections for the following pipes In1, In2, Out1 and Out2 (cf. Figure 6). In the CFD simulation, $C_1 = 0$ mg/L, $C_2 = 1$ mg/L have been chosen. We calculate the ratio θ of the average mass flux Q_3C_3 going out of the output 1 divided by the introduced mass flux:

$$\theta = \frac{Q_3 C_3}{Q_1 C_1 + Q_2 C_2},$$

and in the case $C_1 = 0$ mg/L, $C_2 = 1$ mg/L:

$$\theta = \frac{Q_3 C_3}{Q_2 C_2} = \frac{Q_3 C_3}{Q_2}.$$

We then define θ* as the deviation from the perfect mixing ratio:

$$\theta^* = \theta - \frac{Q_3 \times \frac{(Q_1 C_1 + Q_2 C_2)}{Q_1 + Q_2}}{Q_1 C_1 + Q_2 C_2} = \theta - \frac{Q_3}{Q_1 + Q_2} = \theta - \frac{Q_3}{Q_3 + Q_4}, \quad (22)$$

with θ* being the coefficient saved in the lookup table. θ* is bounded from below by -1 and from above by +1. In practice, θ* was found in the range [-0.056, 0.361].

From $\theta^*(\%\mathrm{ReIn1},\ \%\mathrm{ReIn1},\ \mathrm{Re},\ L/D)$, that interpolated at values not in the table by the Kriging method, $\theta\ (\%\mathrm{ReIn1}, \%\mathrm{ReIn1}, \mathrm{Re}, L/D)$ is computed from Eq. (22).

Therefore, if C_1 = 0 mg/L, C_2 = 1 mg/L:

$$Q_3C_3 = \theta\left(\%In1, \%Out1, Re, \frac{L}{D}\right) \times Q_2.$$

This can be generalized for any C_1 and C_2. The mass flux leaving by outlet Out1 is composed of a part coming from In1 and another one from In2, the same for Out2.

$$\begin{cases} Q_3C_3 = \theta_{13}Q_1C_1 + \theta_{23}Q_2C_2 \\ Q_4C_4 = \theta_{14}Q_1C_1 + \theta_{24}Q_2C_2 \\ \theta_{23} = \theta. \end{cases}$$

We thus have a system of three equations and six unknowns (C_3, C_4, θ_{13}, θ_{23}, θ_{14}, θ_{24}), since we know the flows, C_1 and C_2. To resolve the system, we need three more constraints; they can be found as follows:

When C_1 = 0 mg/L, the mass-balance between inlets and outlets must be satisfied:

$$(Q_1C_1 + Q_2C_2) = Q_2C_2 = (Q_3C_3 + Q_4C_4),$$

so

$$Q_2C_2 = \theta_{23}Q_2C_2 + \theta_{24}Q_2C_2,$$

and assuming that Q_2C_2 is not zero leads to:

$$\theta_{24} = (1 - \theta_{23}).$$

Similarly, when C_2 = 0 mg/L and Q_1C_1 is positive, we get:

$$\theta_{14} = (1 - \theta_{13}).$$

Finally, for C_1=C_2>0 mg/L, the mixing of fluids of similar concentration should give a fluid of same concentration, $C_3 = C_4 = C_1 = C_2 > 0$ mg/L:

$$\begin{cases} Q_3 = \theta_{13}Q_1 + \theta_{23}Q_2 \\ Q_4 = \theta_{14}Q_1 + \theta_{24}Q_2, \end{cases}$$

$$\text{then} \begin{cases} \theta_{13} = \dfrac{Q_3 - \theta_{23}Q_2}{Q_1} \\ \theta_{14} = \dfrac{Q_4 - \theta_{24}Q_2}{Q_1}. \end{cases}$$

This gives four more equations from which three are independent, for example:

$$\left[\theta_{13} = \frac{Q_3 - \theta_{23}Q_2}{Q_1} \;;\; \theta_{24} = (1 - \theta_{23}) \text{ and } \theta_{14} = (1 - \theta_{13})\right] => \quad \theta_{14} = \frac{Q_4 - \theta_{24}Q_2}{Q1}.$$

We have six equations for six unknowns that can be reduced to four equations with four unknowns:

$$\begin{cases} Q_3C_3 = \theta_1Q_1C_1 + \theta_2Q_2C_2 \\ Q_4C_4 = (1 - \theta_1)Q_1C_1 + (1 - \theta_2)Q_2C_2 \\ \theta_2 = \theta \\ \theta_1 = \dfrac{Q_3 - \theta_2Q_2}{Q_1}, \end{cases} \tag{23}$$

with C_3 and C_4 that can be computed from θ, C_1, C_2 and the flow rates.

These equations allow the definition of a law that can be used for every double T-junctions (with equal diameters). CFD simulation results have helped to fill a lookup table with a non-dimensional indicator of the deviation from the perfect mixing. A system of equations that generalizes the case of an introduction of a contaminant only at inlet 2 to both inlets is derived. Firstly, the Kriging interpolation is used to determine the coefficient θ and then is put as input in Eq. (22) to get the concentration at the outlets.

For instance, let us consider a double T-junction with a diameter of 0.1m and an interpipe of length L = 5D. The flow rates are for each input and output $Q_1 = Q_2 = Q_3 = Q_4 =$ 1.96 L/s (the velocity is about 0.25 m/s). Therefore Reynolds numbers are $Re1 = Re2 = Re3 = Re4 = 2{,}500$. The four parameters are respectively, $Re = 5{,}000$, $L = 5D$, $\%In1 = 50$, $\%Out1 = 50$. The interpolation function returns $\theta = 0.6$, then $\theta_2 = 0.6$ and $\theta_1 = 0.4$. Let us say that a contaminated water comes from inlet1 with C_1 = 1 mg/L and another water at inlet 2 with C_2 = 4 mg/L then the concentrations on both outputs are: $C_3 = 2.8$ mg/L and $C_4 = 2.2$ mg/L.

II-7 Two T-junctions Decomposition

The idea is to combine one single T-junction with another T-junction to treat both as separate, henceforth decreasing the computation time. The idea comes from the study of the simulation result of the double T-junction. We can remark that in Figure 24 (top left and right),

the repartition of the scalar inside the intermediate pipe is the same, and at the bottom only the results at the two outlets are different. The entries are dominating the phenomenon. The idea is then to simulate the behaviour in a long T-junction. Then the scalar and the velocities are taken from a cross-section and re-injected as input in another T-junction representing the outputs.

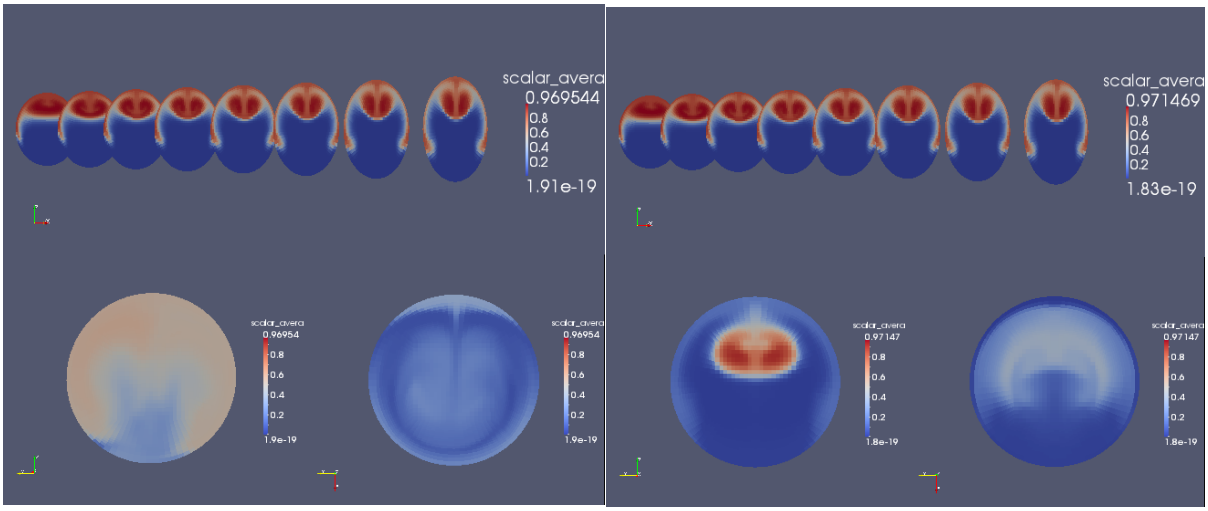


Figure 24: Cross-section in cases 10D, Re=1,000, 80/20 and 80/80.

A study on the use of two T-junctions instead of one double T-junction has been made. The method works on the cases Re = 1,000, L = 5D and L = 10D on most of the points (see Figure 25 and Figure 26). However, the points 70/70 and 70/50 (in 10D case) present a significant error. It works well for Re = 2,000 cases (Figure 27). Finally, for Re = 5,000 (Figure 28) the method shows different results from the double T-junction simulations. In the latter case, more research is needed to conclude on the viability of the method.

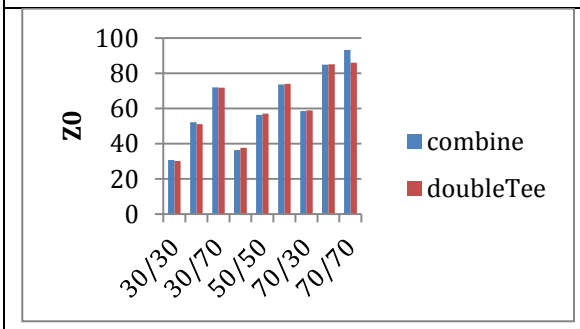
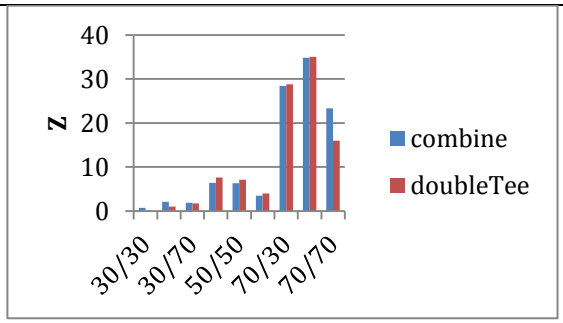
Without projection	Projection with the perfect mixing plan
	

Figure 25: Comparison combined, 5D Re=1,000.

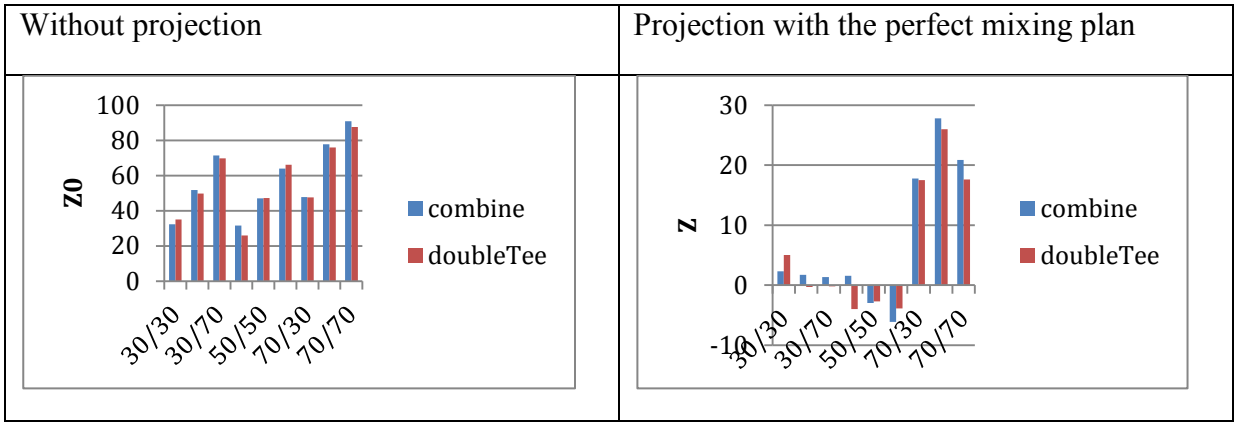


Figure 26: Comparison combined, 10D Re=1,000.

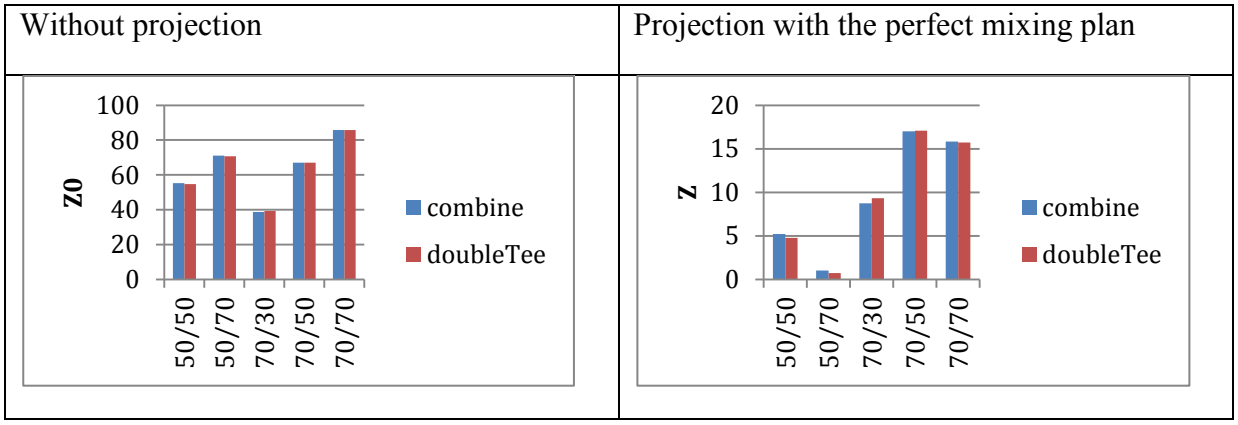


Figure 27: Comparison combined, 5D Re=2,000.

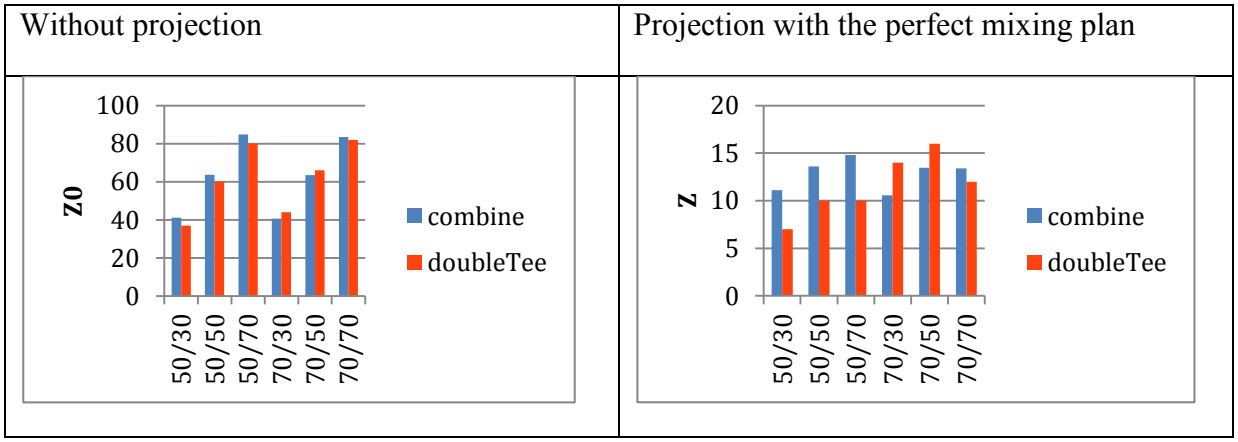


Figure 28: Comparison combined, 5D Re=5,000.

To accelerate the convergence with CFD simulations that requires fewer cells, it was decided to combine two single T-junction simulations to solve the double T-junction problem. It has been found by comparison between full double T-junction simulations and combined single T-junction simulations that this method is working in laminar case for most scenarios. But it does not work for turbulent cases.

II-8 Real Geometry

The real geometry of the experiment case is the one described in Figure 29. Some deviations were observed between CFD simulations and experimentation (cf. Figure 31). It is important to know if the precise modeling of the double T-junction is necessary. It is possible that eddy turbulence phenomenon at elbows may change the mixing ratios at the end. The complex mesh of Figure 30 was generated. It is a 2 million cells mesh generated under Salome software, (see Conn et al. (1988)).

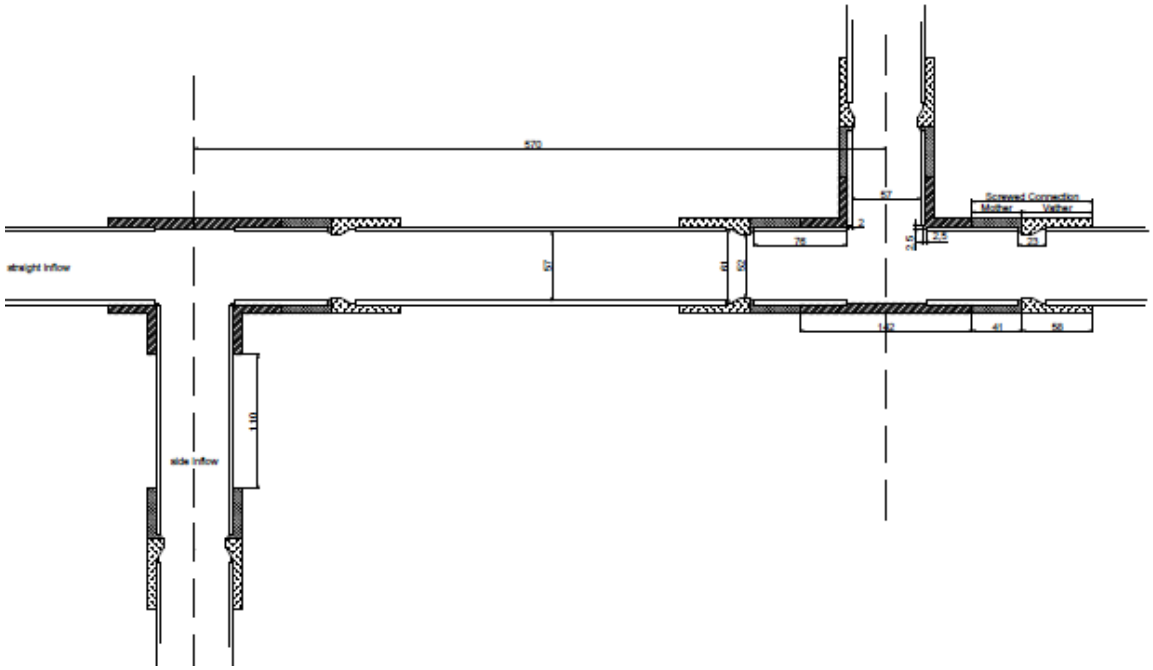


Figure 29: Real geometry.

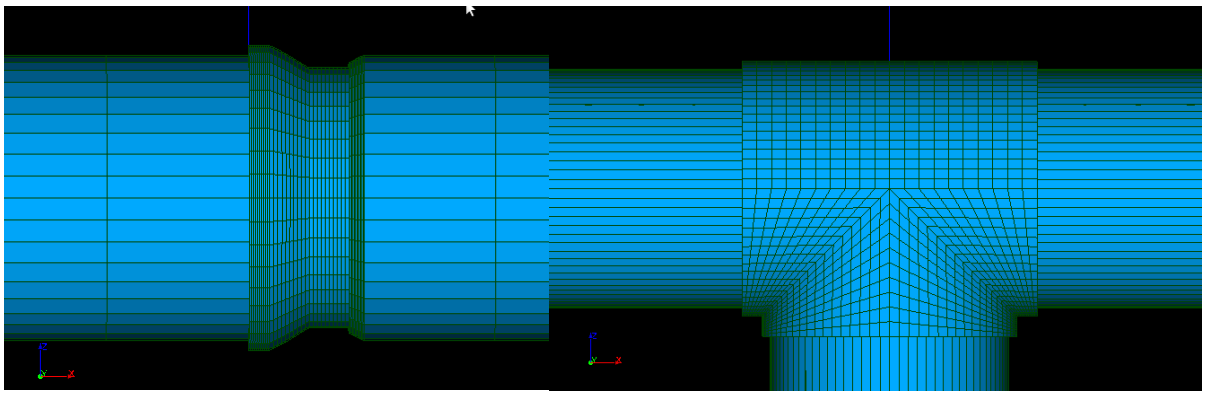


Figure 30: Complex mesh.

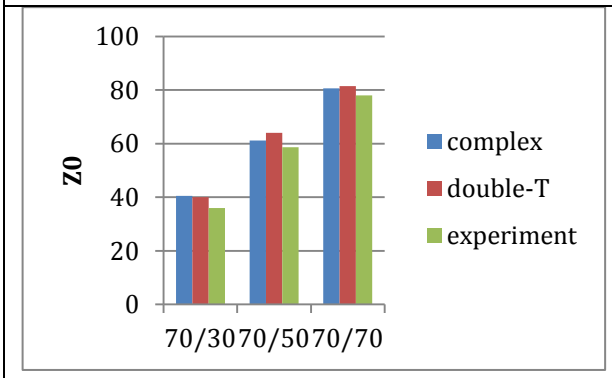
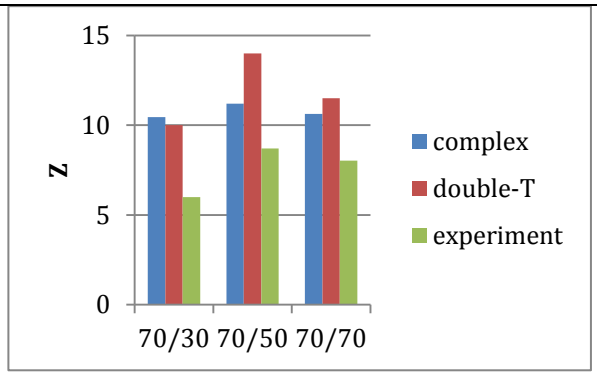
Without projection	Projection with the perfect mixing plan
	

Figure 31: Comparisons of complex geometry, normal geometry and experimentations.

The cases $L = 5D$, $Re = 5{,}000$, $70/(30\,;\ 50\,;\ 70)$ have been simulated but no large deviations can be observed from the previous simulation of double T-junction with simpler mesh.

The complex geometry of the double T-junction caused by joints does not seem to affect the results much. It cannot explain the differences between simulations and experimentations.

II-9 Conclusion

In this chapter, the objective is to improve water distribution quality transport modeling for security management in case of contamination events. The focus is made on mixing conditions at double T-junctions. Numerous 3-D CFD simulations of different configurations of double T-junction mixing were performed and a 1-D law for imperfect mixing was derived that is added to the full network transport model. It depends on the four input parameters: Reynolds number in the interpipe, the length between the two T-junctions and the ratio of Reynolds number at the inlets and at the outlets. A problem-specific Kriging method was developed for interpolating mixing percentages at non-simulated configurations. For sampling design we use a Delaunay method to determine new configurations to simulate.

To complete this study, two CFD simulation tools have been mainly used, Code Saturne for direct numerical simulations (DNSs) and Fluent for Large Eddy Simulation (LESs). The simulations have been launched on the Irstea computational grid and on the MCIA. Additional simulations to test convergence of the grid have been made on the CINES. The results of the simulations have been added to a table and have been interpolated with the interpolation Kriging method.

The Kriging interpolation consists in estimating the function by dividing it in two parts: a stochastic one and a polynomial one. The parameters of the method have been optimized for the problem by minimizing a log-likelihood function. As the problem is not strictly convex and not defined everywhere, choices have been made to simplify the problem. The polynomial coefficients and the variance are taken as their least square optimal values; covariance terms have been chosen as exponential; coefficients in the exponential are functions of $\tilde{\theta}$; and a regularization term α is added to the covariance matrix and chosen as 0.5. Finally, the $\tilde{\theta}$ optimal value is computed to be 3.63 and the order of the polynomial equals to 3. Minimization of the residual variance allows a linear system of equations to be solved.

The computational cost of each simulation justifies the use of a Delaunay method to optimize the repartition of simulation points and therefore the sampling design. The Kriging interpolation is coupled with the Delaunay triangulation method. It consists in calculating interpolation error sums in the triangle domain to define the region the most susceptible to be poorly modeled. It has been compared to a Structured algorithm and performed better with fewer points. It shows that the Delaunay triangulation is a simple and efficient method to find new points of simulation reducing the number of calculations to be made.

The results have been interpreted by fixing both the Reynolds number and the length of the interpipe, or both the Reynolds number ratio at the inputs and at the outputs. The contamination was introduced through one of the two inlets, namely inlet In2. The results show that perfect mixing still occurs when inlet In2 is dominating. It is due to the fluid coming from the input 2 going straight to the wall, creating a lot of turbulence. In the other cases, we observe imperfect mixing with a peak at around $\%\mathrm{In2}=30$. It can also be seen that when the length increases the mixing becomes perfect for all configurations, except in laminar cases where imperfect mixing can still be observed for a length of 20 diameters. In the same way, the increase of the Reynolds number decreases the imperfect mixing impact, but at some point, it converges to a value that may not be the complete mixing.

Finally, a 1-D transport model has been created and implemented. It considers advection/reaction along pipes, imperfect mixing for double T-junctions, and perfect mixing for simple junction nodes. For imperfect mixing the mean cross-sectional concentration at the two outlets is computed. The result depends on the Reynolds number at half distance, on the two inlet and outlet flow rate ratios and on the interpipe length. A lookup table was deduced from the CFD simulations, and a Kriging method is used for points that are not in the table. The imperfect mixing transport model takes the form of a DLL in C++ that can be indifferently called from the Irstea (Porteau) or from the 3S Consult (Sir 3S) software solutions.

This study has been performed for the general case of double T-junctions, composed of pipes with the same diameter, and no slip-wall condition in CFD simulations. Future research will consist in a generalization of the mixing law to pipes of unequal diameters and for pipes with large roughness that influences the turbulence inside the double T-junctions.

CHAPTER III

Enhancing the Transport Model: the Dispersion

Current transport models are designed with only one dimension. To take into account the 3-D phenomena which are the velocity profile and the axial and radial diffusions a new model is necessary. The difference of velocity along the radius induces, for chemical agent, both a front in advance and a back front delay. This is modeled in the literature by a virtual diffusion coefficient added to the transport equation. Different formulations have been proposed that are associated with parameters to be calibrated. This chapter will instead use a backtracking algorithm coupled with a random walk to model the dispersion effect. No calibration is needed but the method is limited to laminar regime with Poisson profile hypothesis. A backtracking scheme is used to focus on the information at the end of the pipe. Then, a random walking scheme is developed to take into account the radial diffusion. Finally, that method is tested on a pilot network.

III-1 Advection-Diffusion Equations

The equations that are mostly used in water distribution network quality model are 1-D advection-reaction equations coupled with perfect mixing at any junction.

$$\frac{\partial C}{\partial t} + u\frac{\partial C}{\partial x} + KC^{\alpha} = 0,$$

with C the concentration of the agent, t the time, x the space, u the average velocity of the pipe and K and α, the kinetic constant and order of the reaction respectively.

$$C_{out} = \frac{\sum_{i \in inflows} Q_i C_i}{\sum_{i \in inflows} Q_i},$$

with C the concentration and Q the flow rate of the pipes whose water is going into the node.

III-1.1 Nondimensionalization

The reaction term is discarded (it can be added after with a splitting scheme), and laminar velocity profile as well as a radial diffusion term are added to the equation in 3 dimensions:

$$\frac{\partial C}{\partial t} + u(r)\frac{\partial C}{\partial x} - D_m \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial C}{\partial r}\right) = 0, \qquad (24)$$

with r the radial position.

In the laminar case,

$$u = 2u_0\left(1 - \frac{r^2}{R^2}\right),$$

then (24) becomes:

$$\frac{\partial C}{\partial t} + 2u_0\left(1 - \frac{r^2}{R^2}\right)\frac{\partial C}{\partial x} - D_m\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial C}{\partial r}\right) = 0, \qquad (25)$$

with the variable changing following:

$$t^* = \frac{u_0}{L}t,$$

$$x^* = \frac{1}{L}x,$$

$$r^* = \frac{1}{R}r,$$

$$C^* = \frac{1}{C_0}C,$$

with C_0 a characteristic concentration not zero.

Then (25) becomes:

$$\frac{u_0}{L}C_0\frac{\partial C^*}{\partial t^*} + 2\frac{u_0}{L}C_0\left(1 - r^{*2}\right)\frac{\partial C^*}{\partial x^*} - D_m\frac{C_0}{R^2}\left(\frac{1}{r^*}\frac{\partial}{\partial r^*}\left(r^*\frac{\partial C^*}{\partial r^*}\right)\right) = 0,$$

and with multiplying by $\frac{L}{u_0C_0}$

$$\frac{\partial C^*}{\partial t^*} + 2\left(1 - r^{*2}\right)\frac{\partial C^*}{\partial x^*} - D_m\frac{L}{u_0R^2}\frac{1}{r^*}\frac{\partial}{\partial r^*}\left(r^*\frac{\partial C^*}{\partial r^*}\right) = 0,$$

also multiplying by ν at top and bottom:

$$\frac{\partial C^*}{\partial t^*} + 2\left(1 - r^{*2}\right)\frac{\partial C^*}{\partial x^*} - \frac{D_m}{\nu}\frac{\nu}{u_0L}\frac{L^2}{R^2}\frac{1}{r^*}\frac{\partial}{\partial r^*}\left(r^*\frac{\partial C^*}{\partial r^*}\right) = 0,$$

therefore:

$$\frac{\partial C^*}{\partial t^*} + 2\left(1 - r^{*2}\right)\frac{\partial C^*}{\partial x^*} - \frac{1}{ScRe}\frac{L^2}{R^2}\frac{1}{r^*}\frac{\partial}{\partial r^*}\left(r^*\frac{\partial C^*}{\partial r^*}\right) = 0, \qquad (26)$$

with Sc the Schmidt number, Re the Reynolds number and because $r^* \in [0,1]$, the radial diffusion term can be neglected only if :

$$\frac{L}{R} \ll \sqrt{ScRe}. \qquad (27)$$

This condition is not always verified, particularly in the night, when the water flows slower than in the day. For example, let us take a laminar case such as $Re = 10^3$, $Sc = 10^3$ and $R = 0.1$ m. Then the condition to neglect radial diffusion is $L \ll 100$ m. This condition is not always respected in hydraulic networks. The radial diffusion cannot be neglected in several cases.

Now if we want to know if the axial diffusion can be neglected, we look for the following nondimensionalized equation:

$$\frac{\partial C}{\partial t} + u\frac{\partial C}{\partial x} - D_m\frac{\partial^2 C}{\partial^2 x} = 0,$$

therefore

$$\frac{\partial C^*}{\partial t^*} + 2(1 - r^{*2})\frac{\partial C^*}{\partial x^*} - \frac{1}{ScRe}\frac{\partial^2 C^*}{\partial x^{*2}} = 0,$$

The criterion to neglect the axial diffusion is:

$$1 \ll ScRe, \qquad (28)$$

which is true in most cases, for instance with $D_m = 10^{-9}\ m^2 s^{-1}$, then $Sc = 10^3$, and therefore the condition is respected for $Re \gg 0.01$. To conclude, the radial diffusion cannot always be neglected but the axial diffusion can be neglected in most cases.

III-1.2 Advective Hypothesis

In order to build an applicable model, we first study the non-diffusive case and integrate the 3 dimensional advection transport. The Poisson profile in a pipe gives the value of the velocity in function of the radius:

$$u(r) = 2u_0\left(1 - \frac{r^2}{R^2}\right), \tag{29}$$

then

$$\frac{1}{S}\int_0^{2\pi}\int_0^{R}\left(\frac{\partial C(t,x,r)}{\partial t}\right)rdrd\theta + \frac{1}{S}\int_0^{2\pi}\int_0^{R}\left(2u_0\left(1-\frac{r^2}{R^2}\right)\frac{\partial C(t,x,r)}{\partial x}\right)rdrd\theta = 0,$$

and by supposing C continuous and differentiable in time and space:

$$\frac{\partial \bar{C}(t,x)}{\partial t} + 2v_m\frac{\partial \bar{C}(t,x)}{\partial x} - \frac{2u_0}{SR^2}\int_0^{2\pi}\int_0^{R}\left(r^2\frac{\partial C(t,x,r)}{\partial x}\right)rdrd\theta = 0,$$

therefore

$$\frac{\partial \bar{C}(t,x)}{\partial t} + u_0\frac{\partial \bar{C}(t,x)}{\partial x} + \xi = 0,$$

with

$$\xi = v_m\frac{\partial \bar{C}(t,x)}{\partial x} - \frac{2u_0}{SR^2}\int_0^{2\pi}\int_0^{R}\left(r^2\frac{\partial C(t,x,r)}{\partial x}\right)rdrd\theta.$$

C is integrated on r between 0 and R:

$$\bar{C}(t,x) = \frac{1}{S}\int_0^{2\pi}\int_0^{R} C(t,x,r)\, rdrd\theta,$$

and from the method of characteristic:

$$C(t,x,r) = C_0\big(T(t,x,r)\big),$$

with

$$T(t,r,x) = t - \frac{x}{u(r)} = t - \frac{x}{2u_0(1-\frac{r^2}{R^2})},$$

and $S = \pi R^2$.

Therefore

$$\bar{C}(t,x) = \frac{2}{R^2}\int_0^{r=R} C_0\left(t - \frac{x}{2u_0\left(1-\frac{r^2}{R^2}\right)}\right) r dr. \tag{30}$$

Let us take the following change of variable:

$$t_1 = t - \frac{x}{2u_0\left(1-\frac{r^2}{R^2}\right)},$$

then

$$r = R\sqrt{1 - \frac{x}{2u_0(t-t_1)}},$$

and

$$dr = R * \frac{1}{2\sqrt{1-\frac{x}{2u_0(t-t_1)}}} \frac{x}{2u_0} \frac{1}{(t-t_1)^2}\, dt_1,$$

by changing those expressions in (30):

$$\bar{C}(t,x) = \frac{x}{2u_0}\int_{-\infty}^{t_1=t-\frac{x}{2u_0}} C_0(t_1)\frac{1}{(t-t_1)^2} dt_1. \tag{31}$$

This gives an explicit relation between $\bar{C}(t,x)$ and $C_0(t_1)$. Let us note that the formula does not depend on the radius of the pipe. In fact, it is numerically computed with the trapezoidal rule. However for some simple cases, as the step case with constant average velocity and laminar profile, it can be solved with an explicit formula.

III.1.3 Step Case

Let us suppose the following contamination profile g(t) which is a step equal to C_0 between the times t1 and t2.

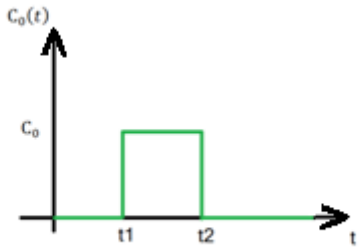


Figure 32: Contamination profile as step definition.

Let us use formula (31), for 3 cases:

$t - \frac{x}{2v_m} < t_1 \Leftrightarrow 2u_0(t - t_1) < x$:

$$\bar{C}(t, x) = 0,$$

$t_1 < t - \frac{x}{2u_0} < t_2 \Leftrightarrow 2u_0(t - t_2) < x < 2u_0(t - t_1)$:

$$\bar{C}(t, x) = C_0\left(1 - \frac{x}{2u_0(t - t_1)}\right),$$

$t_2 < t - \frac{x}{2u_0} \Leftrightarrow x < 2u_0(t - t_2)$:

$$\bar{C}(t, x) = \frac{x}{2u_0} C_0\left(\frac{1}{(t - t_2)} - \frac{1}{(t - t_1)}\right).$$

Therefore we obtain the concentration result with $t_1 = 0s$, $t_2 = 1000s$ and $u_0 = 0.01\ m/s$ plotted in Figure 33.

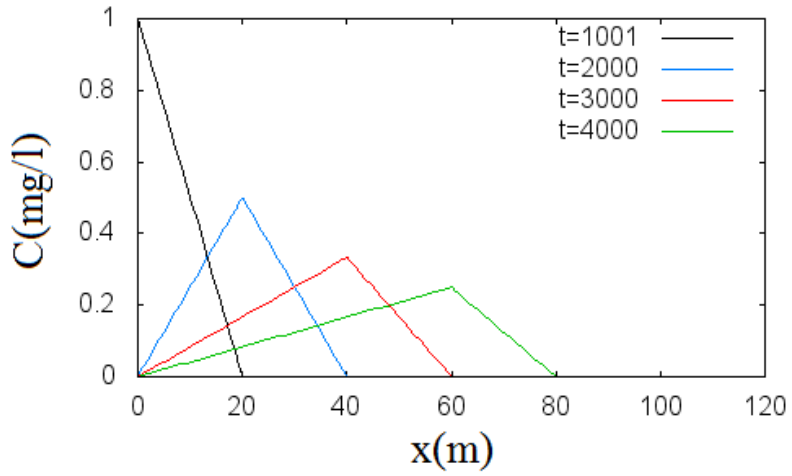


Figure 33: Dispersion effect in pipe along the time.

Now that we have an advective model taking into account 3-D velocity profile, we will then focus in the next part of the case when the radial diffusion cannot be neglected. It will be studied with the use of a backtracking algorithm coupled with a random walk.

III-2 Backtracking Algorithm and Random Walk to Model Radial Diffusion

The backtracking algorithm is a method based on the adjoint formulation of the transport equation on a graph. It is used here to simulate the transport of a chemical agent inside a water distribution network. A random walk algorithm is coupled with it to simulate the radial diffusion.

III-2.1 Equations and Method

The usual transport models used are either Eulerian or Lagrangian. The Lagrangian viewpoint takes advantage of the particle trajectories to workout the solution in backward manner (backtracking algorithm) or forward manner (the forward walking algorithm).

The associated adjoint or inverse transport equations (Neupauer (2011)), when the nonlinear reaction term is not taken into account, permit to go back in time from any node to find where the water comes from.

$$\frac{\partial \psi}{\partial \tau} - u\frac{\partial \psi}{\partial X} = 0, \qquad (32)$$

$$\text{And } \tau = T_0 - t \text{ and } X = X_0 - x,$$

With ψ the adjoint variable, T_0 and X_0 are arbitrary fixed time and length. In the following scheme T_0 is equal to the current time and X_0 the length of the pipe, psi is the adjoint state. Here only advection phenomenon is considered. Water quality indicators with an appropriate change of variable can be described by a simple transport equation (Piller et al. (2014)).

The backtracking algorithm uses the method of the characteristics to find the time needed, for a particle going into one end of the pipe, to get out. The Lagrangian solution takes the form:

$$C(x + dx, t + \frac{dx}{u}) = C(x, t),$$

and its adjoint formulation is:

$$\psi(X - dx, \tau - \frac{dx}{u}) = \psi(X, \tau).$$

If u is constant, and if $dx = L$, then $\frac{dx}{u}$ is the transport time the particle needs to reach the other end of the pipe. The time backtracked t_B is computed using the formula:

$$\int_{t_B}^{t_{end}} u(t) \times dt = \begin{cases} 0 \text{ if reached first} \\ L \text{ else} \end{cases},$$

with t_{end} given and t_B looked for.

This may give multiple solutions. The one chosen here is the first (smallest) non trivial one: $t_B \neq t_{end}$.

That value is computed with the following algorithm. Actually, hydraulic models often use constant velocity by hydraulic time step. In this case, an algorithm has been developed that takes into account that time step Δt_h as explained in the following. First that particle is moved of space step $dx = u(t)\Delta t_h$, with $u(t)$ constant for $t \in [(i-1)\Delta t_h, i\Delta t_h]$ and i an integer representing current time step ($i = 0$ initially). Then a test is made to know if the particle is still inside the pipe. If that is the case, the time backtracked t_B takes the new value $t_B - \Delta t_h$ (with $t_B = t_{end}$ initially) and the algorithm does that first part again with $i = i + 1$, unless the next condition happens. That condition is: the particle is not out and t_B is past the initial time of the hydraulic simulation; then the algorithm ends with t_B set to initial time of hydraulic simulation. However, if the particle goes out, the last time step t_B is only reduced of the part of Δt_h that is necessary for the particle to reach the end of the pipe and then the algorithm ends (see Appendix A-V.8).

III-2.2 Adaptation to Water Quality Transport

The backtracking algorithm is explained for one pipe, but it will be extended to a whole network. Let us take a source of chemical agent N1, a pipe and a consumer N2 and with a concentration of 0 mg/L as initialization for both nodes and pipe:

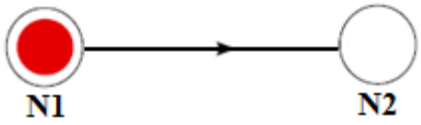


Figure 34: Pipe backtracking.

We want to know the concentration at node N2 for all time under the form $t = i\Delta t_q$, with Δt_q being the quality time step which is a divisor of the hydraulic time step. The backtracking algorithm starts on time t and we look for t_B.

In this example, let us set the velocity to be constant, $u = 0.1$ m/s, the length of the pipe to be equal to $L = 100$ meters and the source N1 to have a constant concentration C_1 = 1 mg/L from time t = 0 s (0 mg/L before). The quality time step is fixed to 300 seconds.

First a backtracking is used at time $t = 300s$ at N2, the output is N1 at time $t_B = -700s$, because the time needed to go through the entire pipe is u×L which is equal to 1000s. Therefore, thanks to the method of the characteristics, we know that the concentration at N2 at 300s is the same as the one at N1 at -700s: $C_2(300s) = 0$ mg/L. In the same way, C_2 gets the same value for time $t = 600s$ and $t = 900s$. For starting times above 1200s the results of the backtracking are times $t_B > 0s$. As $C_1(t > 0s) = 1$ mg/L, therefore, in this example case, $C_2(t \geq 1200s) = 1$ mg/L.

The algorithm can be extended to a graph. We take one source of chemical agent N1, four pipes, two passing nodes N2, N3, a consumer N4 and a concentration of 0 mg/L as initialization for all nodes and pipes:

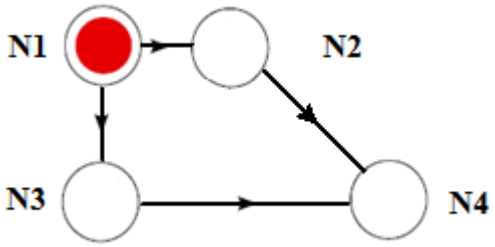


Figure 35: Small network backtracking.

We take the following example: all pipes have equal flow rate and it takes 300s to go from N1 to N2, 500s from N1 to N3, 600s from N2 to N4 and 800s from N3 to N4. In Table 6 the concentration values of all nodes for the 5 first time steps are summarized.

Table 6: Transport for small network.

t (s)	C_1 (mg/L)	C_2 (mg/L)	C_3 (mg/L)	C_4 (mg/L)
< 0	0	0	0	0
0	1	0	0	0
300	1	1	0	0
600	1	1	1	0
900	1	1	1	0.50
1200	1	1	1	0.66
1500	1	1	1	1

The concentration of C_2 and C_3 are computed with the method presented previously on one pipe. The concentration of C_4 is calculated as follows:

$$C_4(T_0) = \frac{Q_2 C_2(T_0 - t_{2\to4}) + Q_3 C_3(T_0 - t_{3\to4})}{Q_2 + Q_3} = \frac{C_2(T_0 - t_{2\to4}) + C_3(T_0 - t_{3\to4}))}{2},$$

with $t_{i\to j}$ the time it takes to go from node Ni to node Nj, Qi the flow rates and Ci the concentrations.

A linear interpolation is used on the concentration for any time backtracked t_B that is not a multiple of the quality time step. Let n be the integer that respects

$$t_B \in [t_1 = (n-1)\Delta t_q, t_2 = n\Delta t_q],$$

then:

$$C_{out}(t) = \frac{t - t_1}{t_2 - t_1} C_{in}(t_2) + C_{in}(t_1),$$

and therefore in the example case:

$$C_4(1200) = \frac{C_2(600) + C_3(400)}{2} = \frac{1 + 0.33}{2} = 0.66 \text{ mg/L}.$$

III-2.3 Random Walk Coupled to the Backtracking Algorithm

The backtracking algorithm, modeling the axial transport, is coupled with a random walk method to model the radial diffusion effect. Let us consider a pipe of length L and a laminar velocity profile. A concentration C(t) is injected in the inlet, and depends only on the time (uniform on the cross-section). We want to find the concentration C(x,r,t) to be integrated on the cross section at x.

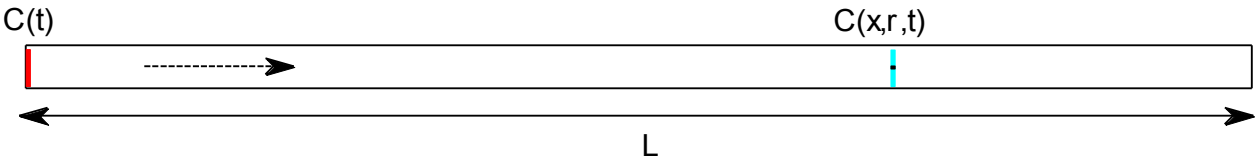


Figure 36: Pipe, dispersion problem.

The direct method consists in injecting particles along the cross-section at the inlet of the pipe to count how many reach x. This implies a large number of particles in time and space. The idea is to use the particle backtracking (non reactive tracer). Taking the end of the pipe, particles are injected and they are moved back in time and space to find their original position (radius) and the time they have been injected in order to find the concentration they had at that time.

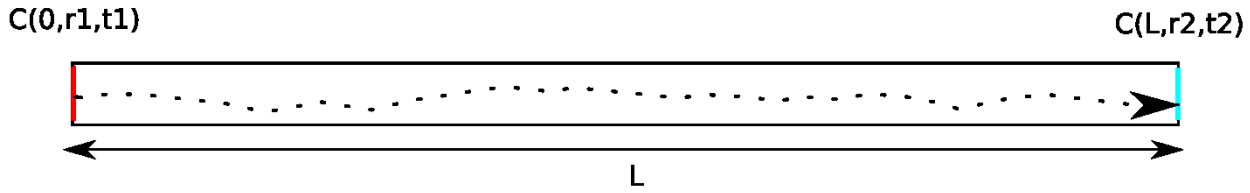


Figure 37: Pipe, particle backtracking with random walk.

The advantage of the backtracking is that the information of the introduced particles is where we want to know the results, at at the junction end of the pipe. Therefore, fewer particles are needed to get the result.

As it is the application of the analytical formula (Eq. 31) computed with a rectangle rule, the results are the same.

To simulate the radial diffusion, a random walk method has been used: where each backward movement on x and time step dt is associated to a radial movement with quadratic average:

$$< X^2(t) > = d \times D_m \times dt,$$

with d the spatial dimension, D_m the coefficient of diffusion and dt the time step.

A Gaussian can be used to determine the amplitude of the random walk. Therefore, in one dimension, the move of one particle is given by:

$$dx = G(x, D_m \times dt).$$

In a pipe, the particle is moving in a disk corresponding to the cross-section. Both movement in x and y direction are given by a Gaussian.

$$dx = G(x, D_m \times dt),$$

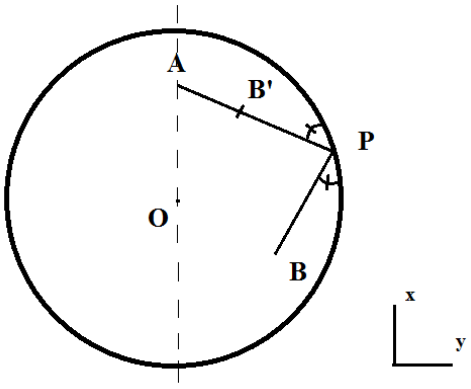
$$dy = G(x, D_m \times dt),$$

which have been implemented with the current formula:

$$G(x, D_m \times dt) = \sqrt{2D_m \times dt} \times \sqrt{-2\ln(\eta_1)}\cos(2\pi \times \eta_2),$$

with η_1 and η_2 two random numbers between 0 and 1.

The boundaries are taken as reflective. A is the starting point, B the ending point.



1

Figure 38: Random walk with circle boundary condition.

As the problem is symmetric, dy is taken positive to ease the calculation. P is the point that intersects the line (AP) and the circle.

$$\begin{cases} x^2 + y^2 = R \\ x - \dfrac{dx}{dy} y = r, \end{cases}$$

with r the position of A in the x-axis and R the radius of the circle. By substitution and by resolving a second-order polynomial the coordinates of the point P are computed:

$$\begin{cases} y = \dfrac{-r\dfrac{dx}{dy} + \sqrt{R^2 - r^2 + \dfrac{dx}{dy}^2 R^2}}{1 + \dfrac{dx}{dy}^2} \\ x = r + \dfrac{dx}{dy} y, \end{cases}$$

with B' being symmetric to the point B around (OP) and therefore it has the same radius. B' is calculated from P with the vector (-dx;-dy) and the amplitude $\left|\overrightarrow{PB}\right| = |(dx; dy)| - \left|\overrightarrow{AP}\right|$.

III-2.4 Random Walk Validation

Different methods have been described that solves the transport problem:

- The Eulerian hybrid method of Porteau (see I-II.1) that will be called the no dispersion model.

- The backtracking in one dimension (see III-2.2)

- The integral formula (Eq. (31))

- The backtracking and random walk with no radial diffusion which give the same result as the integral formula in constant velocity (see III-2.3 with $D_m = 0\ m^2/s$) that will be called the no radial diffusion model.

- The backtracking and random walk with a radial diffusion D_m (see III-2.3)

Here, the results of those models are compared to CFD simulations results. Pipes of 1 and 2 meters, with a diameter of 100 mm, are meshed and a laminar model is used to simulate the propagation of contamination with a concentration equal to 1 mg/L on the whole cross section at the inlet. The advection-diffusion equation is simulated with diffusive coefficient $D_m = 10^{-6}\ m^2/s$ and $D_m = 10^{-5}\ m^2/s$.

The simulations have been performed with Code Saturne software, (see Archambeau et al. (2015)). The pipe is meshed as illustrated in Figure 39 with 30 cells on the radius (20 in the boundary layer) and 320 on the length. Only one quarter of the pipe section is used and symmetric boundary conditions are applied on the surface cut. Dirichlet condition with zero velocity is used on the round surface. Laminar velocity profile is imposed on the entry and at the initialization inside the pipe.

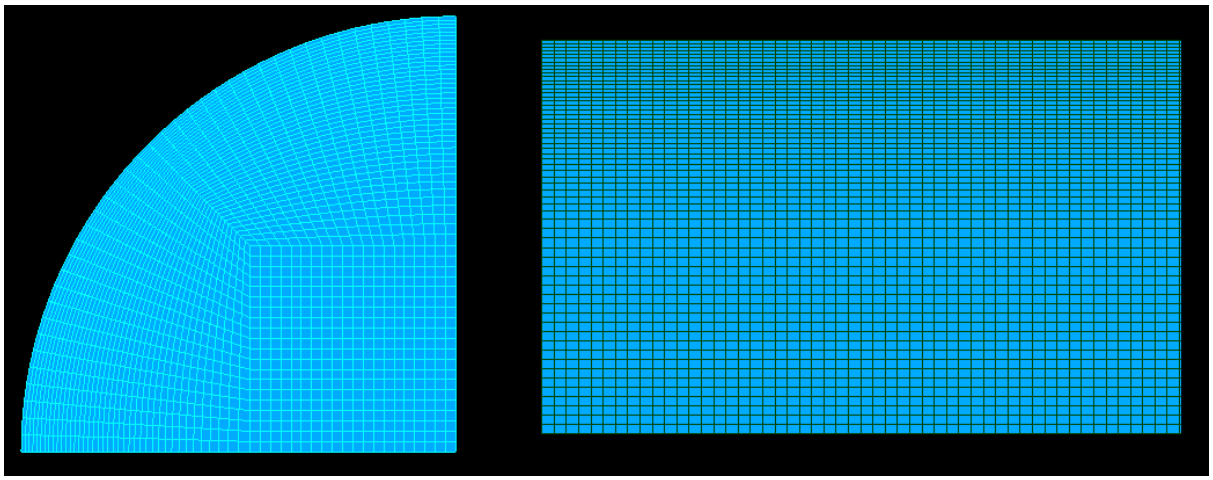


Figure 39: Pipe mesh.

The Figure 40 shows a comparison between the different models, considering the simulation of the propagation of a contamination in a 1 meter pipe. The purple line is the model without dispersion and the green line corresponds to the results when only taking into account the velocity profile effect (no radial diffusion). The green one is reaching the end of the pipe twice faster than the purple one. The CFD result gives a curve in-between those two curves; it is due to the radial diffusion that tends to make uniform the concentration in the cross section. Finally, the backtracking and random walk model with $D_m = 10^{-6}\ m^2/s$ (blue line hidden by red line) give similar results with the CFD results for both 1 meter and 2 meters.

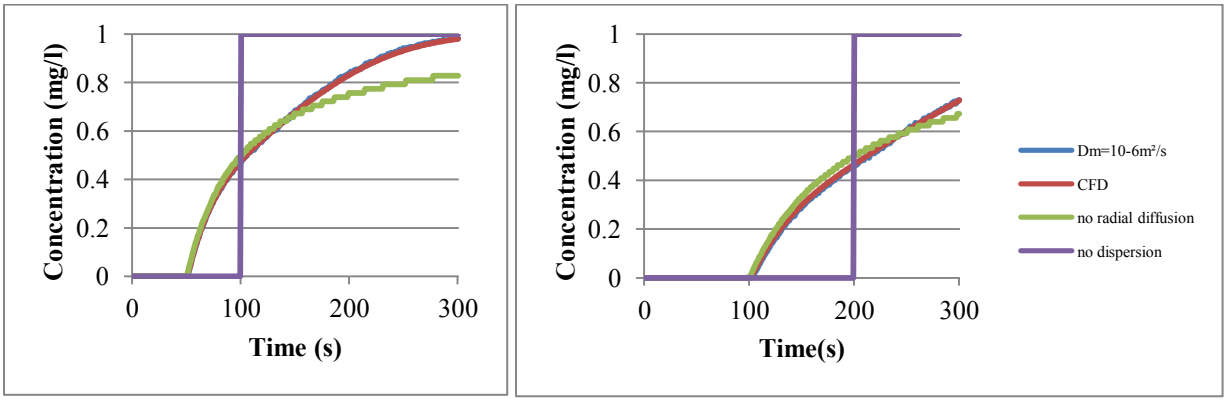


Figure 40: Comparison CFD, dispersion $D_m = 10^{-6}\ m^2/s$ with and without radial diffusion and no dispersion on a 1 and 2 meter pipe.

The same comparison has been made with the diffusion coefficient $D_m = 10^{-5}\ m^2/s$. As shown in Figure 41 the backtracking and random walk model with $D_m = 10^{-5}\ m^2/s$ also give results similar to the CFD results for the 1 meter pipe. This proves the efficiency of the backtracking and random walk model for different diffusive coefficients and different lengths. It needs several hours on 12 processors to do the CFD simulation and only several minutes on 1 processor to do the same with the backtracking and random walk model.

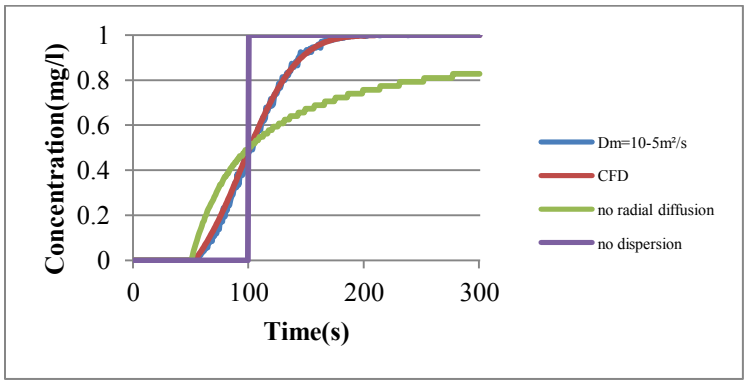


Figure 41: Comparison CFD, dispersion $D_m = 10^{-6}\ m^2/s$ with and without radial diffusion and no dispersion on a 1 meter pipe.

The results for the different models are compared in Figure 42. As shown earlier, the no dispersion model and the no radial diffusion model give differen²t results. When taking into account the radial diffusion, the result is in-between those two models. The bigger is the diffusive coefficient, the closer it is to the no dispersion model.

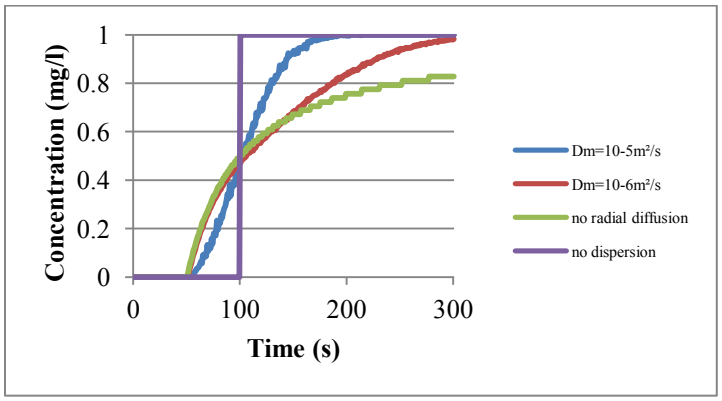


Figure 42: Comparison dispersion $D_m = 10^{-6}\ m^2/s, D_m = 10^{-5}\ m^2/s$ with and without radial diffusion and no dispersion on a 1 meter pipe.

The result is also given for a step shape concentration at inlet in Figure 43.

Diffusive coefficients below $D_m = 10^{-7}\ m^2/s$ give results close to the no radial velocity model. The concentration arrives twice faster than the no dispersion model. The curve increases progressively until a certain time, then decreases slowly. That time corresponds to half the time it takes for the end of the step to reach the end of the pipe with the no dispersion model.

A Gaussian shape can be seen for diffusive coefficients higher than $D_m = 10^{-6}\ m^2/s$ and therefore may be modeled by a virtual diffusion term as it has been treated in literature, see Taylor (1953), Romero-Gomez et al. (2009).

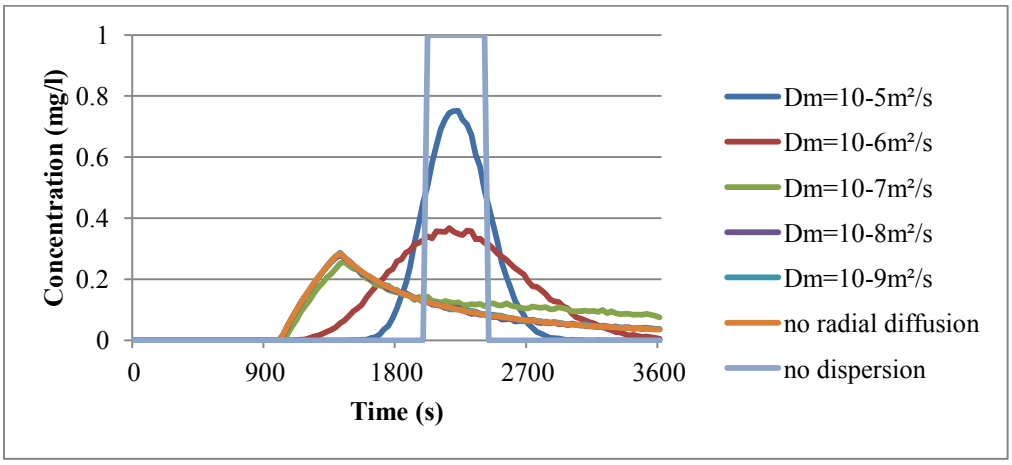


Figure 43: Comparison dispersion $D_m = 10^{-9}\ m^2/s$ to $D_m = 10^{-5}\ m^2/s$ with and without radial diffusion and no dispersion on a 20 meter pipe with step concentration.

III-3 Application to a Small Network and an Experiment Pilot

The algorithm coupling backtracking to a random walk is compared here to no dispersion models. The following transport models are compared: Eulerian hybrid transport model implemented in Porteau, the backtracking algorithm in one dimension, the no radial diffusion algorithm, and two cases of backtracking with random walk that have diffusive coefficient $D_m = 10^{-9}\ m^2/s$ and $D_m = 10^{-6}\ m^2/s$ respectively. It is then applied to a pilot scale (see Figure 46) for discussion.

The network is presented in Figure 44, it is composed of two pipes, one source "rs0" and one consumer "od2". A periodic step concentration is injected in "rs0" and the profile of consumption, and therefore the velocities in the pipe, is assumed to be oscillating. The Reynolds number is oscillating between 250 and 3750.

The Poiseuille profile for $Re \leq 2000$ is defined as follows:

$$u(r) = 2u_0\left(1 - \frac{r^2}{R^2}\right).$$

For high Reynolds number ($Re \geq 4000$) the profile is taken as the average axial velocity:

$$u(r) = u_{0,}$$

and finally, to take into account the transitional regime, Reynolds number in-between, a linear prolongation of the velocity profile is taken as followed:

$$u(r) = \frac{Re - 2000}{2000}\left(u_0 - 2u_0\left(1 - \frac{r^2}{R^2}\right)\right) + 2u_0\left(1 - \frac{r^2}{R^2}\right). \quad (33)$$

III.3.1 Model Comparison

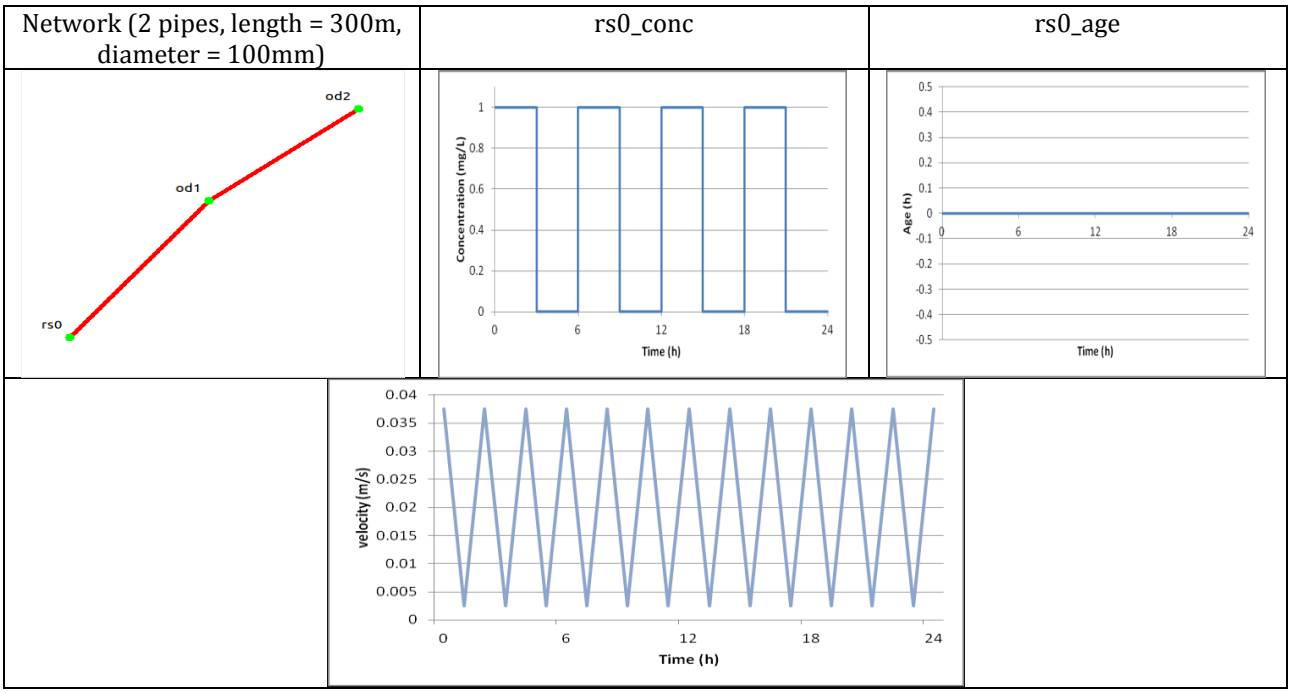


Figure 44: Network, concentration and age at "rs0", velocity in pipes rs0->od1 and od1->od2.

The Figure 45 shows the results (concentrations evolution with time at node od1 and od2) given by the different models. The no dispersion model gives a step. The no radial diffusion model achieves to a different solution. In the two situations $D_m = 10^{-9}\ m^2/s$ and $D_m = 10^{-8}\ m^2/s$ associated to the backtracked with random walk model, because of their low radial diffusivity, have results close to the no radial diffusion model. Logically, the other diffusive coefficients associated to the backtracked with random walk model show a concentration between no radial diffusion and non dispersive cases.

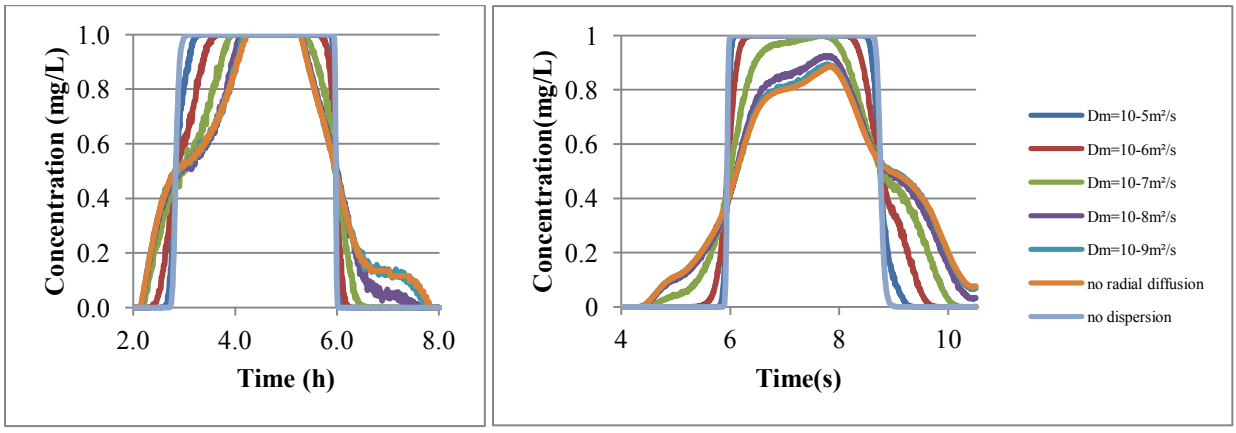


Figure 45: Comparison concentration at node od1 and od2 for models with dispersion $D_m = 10^{-9}\ m^2/s$ to $D_m = 10^{-5}\ m^2/s$ with and without radial diffusion and no dispersion on Network.

Table 7 gives the computer running times for these models with a time step of 30 seconds and the simulation time fixed to 12h. The dispersion models have used 1000 particles. Depending on the targeted accuracy, less or more will be needed. The no radial diffusion model takes fifteen more time than the no dispersion model, so it may still be used for calculation. However, the full dispersion models with radial diffusion takes until 400 times more time. They can be used with bigger quality time steps, but only for small networks.

Table 7: Model execution times comparison with dt = 30s.

D_m	10^{-5} m^2s^{-1}	10^{-6} m^2s^{-1}	10^{-7} m^2s^{-1}	10^{-8} m^2s^{-1}	10^{-9} m^2s^{-1}	No radial diffusion	No dispersion
Running time	432	400	377	370	360	15	1

III-3.4 Experiments

Experiments have been performed on the TZW network in Dresden to calibrate the pilot network with measures of conductivity of a chemical product injected inside the network.

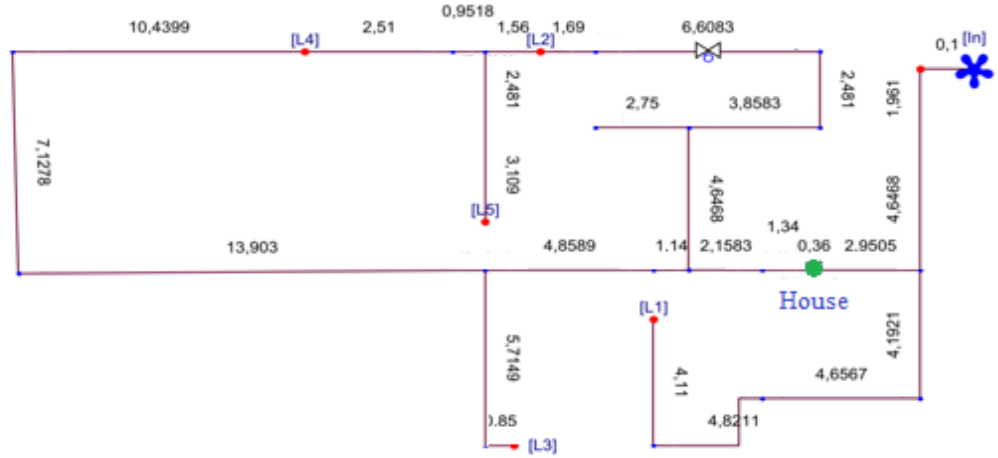


Figure 46: Pilot network TZW.

In Figure 46 the pilot network is represented with the length of the pipe in meters. The conductivity sensors are placed at red points “L1”, “L2”, “L3”, “L4” and “L5” (red nodes) with the injection of contaminant (constant profile from 60s during 120s) at point “House” (green node). The hydraulic paths go from resource node “In” to the points where the sensors are places.

Moreover velocity sensors have been placed on different pipes and the other pipes velocities are computed with the help of the mass conservation law. The measures of conductivity have been adapted in order to be compared with simulation results, by equalizing the mean integral of the 4 sensors to the concentration introduced at the inlet which have been used for the simulation.

The concentration at each sensor node is calculated with Porteau. Figure 47, Figure 48 and Figure 49 show results for three different hydraulic permanent regimes 1, 2 and 3. The first one is in laminar regime (Re < 2000), the second in the transitional/turbulent (2000 < Re < 6000) state and the last is in turbulent regime (Re > 4000). For Porteau simulations, a step has been injected, therefore, in such 1-D simulations, it is normal to find steps at outputs.

Figure 47, Figure 48 and Figure 49 show the comparison between Sensors L3 and L5 measurements and the 3 models: no dispersion, no radial diffusion and full dispersion with diffusive coefficient $D_m = 10^{-6}\ m^2/s$. These figures present respectively the results found in regimes 1, 2 and 3.

For laminar regime 1 in Figure 47, the curve that corresponds to the new backtracking and random walk model, with diffusive coefficient $D_m = 10^{-6}\ m^2/s$, gives satisfactory results compared with experiments. The no dispersion model is late on the measurements and is overestimating the concentration. On the contrary, the no radial diffusion model estimates a concentration lower than in reality, and also does not foresee the two fronts of concentration for the sensor L5.

The transitional/turbulent regime 2 is presented in Figure 48. The dispersion model still does well for the transitional case. The model without dispersion is closer to experiments than in the previous regime. Finally, the Figure 49 presents the turbulent regime 3. The model is less good as it has not been adapted to the turbulent case in this thesis.

In all cases and models, the measurements are in advance from what the models predict. It can only be due to a real phenomenon that is not taken into account. That can be the value of the velocities estimated from the conductivity measurements, that are superior to those measured by the velocity sensors. It can also be due to the fact that conductivity meters are placed at the center of the pipe and therefore close to high velocities.

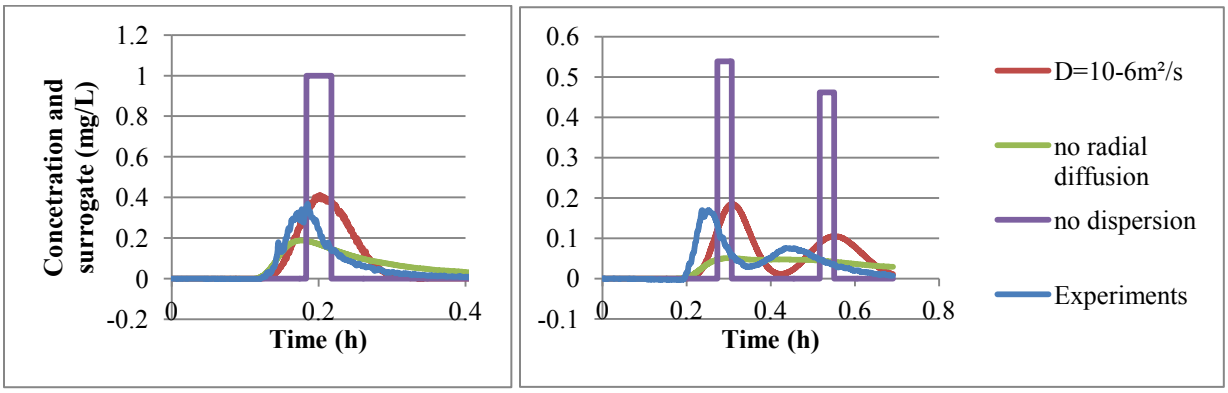


Figure 47: Comparison at Sensor L3 and L5, regime 1.

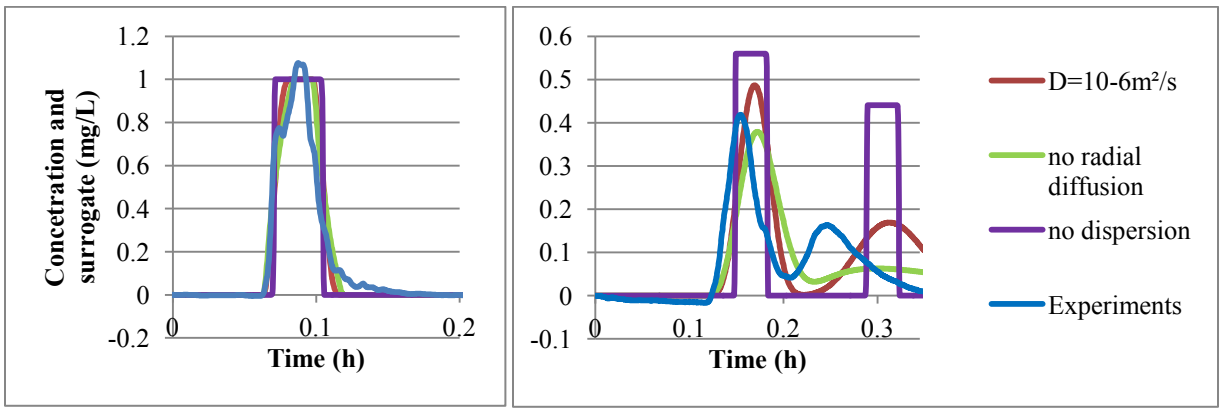


Figure 48: Comparison at Sensor L3 and L5, regime 2.

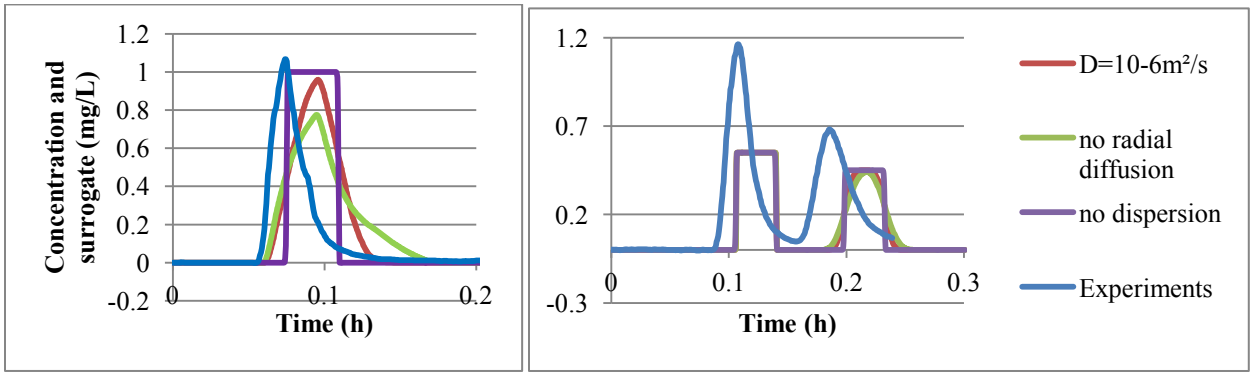


Figure 49: Comparison at Sensor L3 and L5, regime 3.

III-4 Conclusion

This chapter tackles the dispersion effect in water distribution systems transport models. This consists in taking into account two phenomena which are the velocity profile and the radial diffusion in the pipe. It has been modeled in different papers by a virtual diffusion coefficient using various different models and it needs to be calibrated. The approach here is a backtracking scheme coupled with a random walk. The first one models the velocity profile, the second the radial diffusion. It does not need calibration however it is highly dependent on the velocity and is only treated in laminar cases in this chapter.

The nondimensionalization of the advection diffusion equation in laminar regimes has given two criteria to neglect the radial diffusion and the axial diffusion. The radial diffusion term can be neglected only if $\frac{L}{R} \ll \sqrt{ScRe}$ and the axial diffusion can be neglected if $1 \ll ScRe$. The first one is hard to get in water distribution models but the second can easily be found. If the two conditions are respected, diffusion can be neglected, and an integral formula can be given to calculate the concentration anywhere inside a pipe, knowing the concentration limit conditions at the ends of the pipe:

$$\bar{C}(t,x) = \frac{x}{2v_m}\int_{-\infty}^{t_1=t-\frac{x}{2v_m}} C_0(t_1)\frac{1}{(t_1-t)^2}dt_1.$$

Therefore, a model of backtracking has been developed for water distribution networks. It uses the adjoint formula of the transport equation to link the concentration at one end of the pipe to its origin and deduce its value from it. It has been extended to more than one pipe using a concentration time table at all nodes. A discretization of the radius on the cross section of the pipe permits to take into account the velocity profile and gives the same results as the integral formula. Finally, a random walk is added to the scheme to model the radial diffusion. It has been compared to CFD results and gives good results for different diffusion coefficients.

The model has been used to compare the results of the different transport schemes on a double pipe network with oscillating velocity and a periodic step concentration. When taking into account the dispersion effect, the results for concentration are different than one dimensional with no dispersion effect models. The hydrodynamic diffusion (with only velocity profile) can be modeled for a cost of fifteen times the usual transport calculation time, with a discretization $N = 100$ which has been sufficient in the case tested. Moreover, the radial diffusion model by random walk is expensive in time, around 400 times the usual time. It depends greatly on the quality time step used as well as the diffusive coefficient. Research in parallelization can be made to accelerate the calculations.

Finally the method has been tested to model a pilot scale where current model simulation results are different than those measured. Three regimes have been tested, laminar, transitional/turbulent and turbulent. The backtracking scheme and random walk models have generated results close to measurements in both first and second experiment cases, but less good in turbulent case. More research is needed to take into account the dispersion in turbulent regime.

CHAPTER IV

Contamination Source Identification

The enhancement of the transport model is needed to precisely estimate the propagation of a contaminant. In this chapter, we tackle the problem of source identification. It is solved with the use of a backtracking algorithm to calculate the list of potential source nodes. It also gives a ranking of those nodes to be the real source of contamination.

Real-time sensors are quite a new topic. Multiple researches are focused on where to place the sensors and how to treat the data they produce. One possible use is the identification of the source of a contamination. Taking into account that we have an alarm generator based on non specific quality sensor, the aim is to find the location of the source of the contamination by using the time and location of detection as well as the history of the network dynamics.

We consider binary sensor responses: positive or negative. Both can be used to deduce the location, starting time and duration of a contamination. A positive response means that the alarm generator algorithm has computed that, at this sensor location and at that time, there has been a contaminated water going through. In the other case, a negative response is produced that also helps in identifying the node where no contamination could have happened. In this thesis, only perfect sensors are considered: there is no false positive and false positive.

This chapter first presents the source identification method used as well as the backtracking algorithm. Further, we first describe the multi-contamination solution and second the criteria used for the sensor placements.

IV-1 Backtracking Algorithm

The backtracking algorithm is a method based on the adjoint formulation of the transport equation on a graph. It has been used by Shang et al. (2002) and De Sanctis et al. (2010) to enumerate the potential sources of contamination. It is used here to build the input/output matrix as defined in Propato et al. (2010) but beginning at positive sensor responses in order to accelerate the calculations. The negative responses are also used as presented afterwards but not as part of the matrix of contamination. That matrix is a binary matrix, where true values link the potential source of contamination to positive sensor responses. Negative responses help identifying times where nodes can't be source of contamination and therefore limit true values in the matrix.

IV-1.1 Equations and Method

As seen in part III-2.1, the equations that are mostly used in water distribution network quality models are one-dimension advection-reaction equations coupled with perfect mixing at all junctions.

$$\frac{\partial C}{\partial t} + u\frac{\partial C}{\partial x} + KC^{\alpha} = 0,$$

with C the concentration of the agent, t the time, x the space, u the average velocity of the pipe and K and α, the kinetic constant and order of the reaction respectively.

$$C_{out} = \frac{\sum_{i \in entries} Q_i C_i}{\sum_{i \in entries} Q_i},$$

with C the concentration and Q the flow rate.

The contaminant goes from a source, travels through the network and may reach an installed sensor. The usual transport models used are either Eulerian or Lagrangian. The adjoint equations associated allow to go back in time from the sensor response to find in the network the possible sources.

$$\frac{\partial \psi}{\partial \tau} - u\frac{\partial \psi}{\partial \chi} = 0,$$

$$\text{And } \tau = T_0 - t \text{ and } \chi = X_0 - x.$$

The backtracking algorithm used is one of the Lagrangian adjoint forms. It uses the method of characteristics to find the time needed, for a particle going into one end of the pipe, to exit out. The Lagrangian solution takes the form:

$$C(x + dx, t + \frac{dx}{u}) = C(x, t),$$

and its adjoint formulation is:

$$\psi(X - dx, \tau - \frac{dx}{u}) = \psi(X, \tau).$$

If u is constant, and if dx = L, then $\frac{dx}{u}$ is the time it needs to reach the other end of the pipe. Hydraulic models often use constant velocity by hydraulic time step. In this case the time is obtained with the formula:

$$\int_{T_B}^{T_{end}} u(t)dt = \begin{cases} 0 \text{ if reached first} \\ L \text{ else} \end{cases},$$

with T_{end} given (the time step calculated) and T_B looked for.

This may give multiple solutions, the one chosen is the first (smallest) that is not the trivial one ($T_B = T_{end}$).

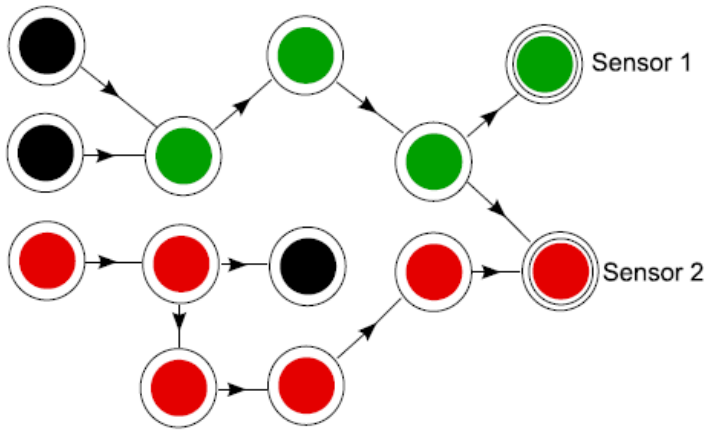


Figure 50: Backtracking algorithm method, red are potential nodes, green are safe nodes and black are unknown nodes.

To get the whole solution around the network, a recursive algorithm is used as shown in Figure 50. The red sensor has detected a contamination as contrary to the green one. Green nodes are safe and red ones are potential sources of contamination. Top nodes are safe because green wins upon red. Indeed, if the top nodes were red, the Sensor 1 would have been red. Finally, black nodes are the ones upon where we don't have any information. The middle one is black because no sensor monitors it. The two nodes on the top left are black because the time limit for the backtracking has been reached.

Let us take a contamination detected at a node at a time. We calculate for every inflow pipe the node and the time from when the contamination could have originated and repeat the algorithm for each new node reached. Moreover, as this method can go back far in time, a backtracking time (BT) is given from when it is assumed that the contamination cannot have lasted that much time, it needs to be calibrated from the maximum residence time inside the network. Also, as will be shown in the results part, the efficiency of the source identification method depends on both BT and the average time to detection.

This algorithm allows from a sensor node and a response time to calculate every node location and time that can explain where this detection is coming from. Now, it can also be used in another way, if the sensor is sure to not have detected something, the algorithm can tell which node at which time cannot be a source of contamination.

Finally, both ways can be used simultaneously and for a large range of sensor times by launching all simulations and crossing the information. This allows the creation of the input/output matrix of contamination. It is the matrix that can illustrate which node at which time can explain which positive sensor at which time. Henceforth, it gives the list of potential sources of contamination. An example is given in Figure 51.

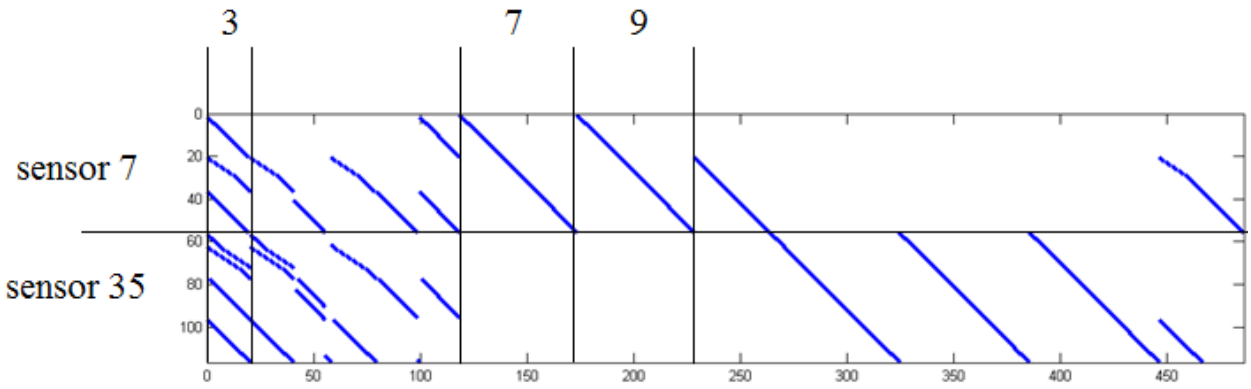


Figure 51: Input/output matrix of contamination, blue points are non zero values.

In Figure 51, each row represents a positive sensor response at a time and the column a potential node at a time. A blue point means that the node at a time corresponding to a column can explain the sensor node at a time on the row. Each sub-diagonal matrix represents the contamination duration. Indeed, in case of a velocity that does not vary a lot, subsequent contamination time steps can explain subsequent sensor responses.

Now, by using the input/output matrix of contamination, it is possible to find which node may be more likely to be the real source of contamination. Indeed, the node that best explains the sensor response is the one with the most non zero values in every row, all times taken together.

Table 8: Simplified input/output contamination matrix.

Node Id	3	5	7	9	17	19	21	23	31	33	35
Sensor 7	0.98	0.69	1.00	1.00	0.56	0.64	0.64	0.64	0.00	0.00	0.00
Sensor 35	0.98	0.30	0.00	0.00	0.98	0.89	0.33	0.00	1.00	1.00	1.00
Mean	**0.98**	**0.49**	**0.50**	**0.50**	**0.77**	**0.76**	**0.48**	**0.32**	**0.50**	**0.50**	**0.50**
Rank	1	9	8	8	2	3	10	11	8	8	8

Table 8 shows the results for simplification (aggregation) in time of the matrix in Figure 51, on the rows and the columns. Firstly, columns of the same node are combined in one column that takes the value 1 on each row having at least one non zero value. Then all columns are averaged on the number of rows to gives a score. Table 8 gives the average score for each sensor. In fact, all sensors are weighted with their number of rows, or positive responses, but in this example both sensors have the same weight. The averaging on the rows gives an efficient weighting for ranking nodes to be the likely source of contamination. In this example, node 3 is the one that is most likely to be the true source of contamination. It can explain 98% of the positive responses at sensors. However the mean taken is calculated on all responses, no matter the sensor it belongs to, and therefore the method favors the information given by the sensors with more responses.

The averaging method used is algebraic, instead, a product may have been used to define probabilities. Indeed, in case of one contamination only, node 7 in Table 8 has a zero probability to be the source of contamination if the product of the scores in the rows of sensor 7 and 35 score is done. The method used here is more robust, all positive sensor responses have the same weight. Also it can be used in case of multi-contamination contrary to the product probabilities, as explained in the next part.

Two variables of the backtracking algorithm are defined: the backtracking time (BT) and the observation time (OT) from first time to detection. These two variables define two real aspects of source identification. The first one is the time the algorithm goes back in time to find the source of contamination. The second one is the period of observation used after the first observation. The ability of a sensor placement to do the source identification will depend on these two parameters. A large BT will allow finding contamination that happened far in the past and a large OT means more information to solve the inverse problem. A good sensor placement tries to minimize these two parameters while keeping the same results. A small BT induces a good early warning, the contamination is detected fast. A small OT proves a good performance; the amount of observation time needed to accurately identify the contamination is small.

IV-1.2 Multi-Contamination

The problem of source identification with multi-contamination can also be tackled. It is solved here with a simple application of coverage theory (see Godsil et al. (2004)). The method that has been presented, as such, cannot establish if it is a one or multi-contamination case. However by operating on the matrix it is possible to define cross information. Generally, a product of the matrix with its transpose is used to create the matrix of information. However, the simplified input/output matrix of contamination (before row averaging) is binary and therefore the operator OR is used here. This operator allows the definition of couples of nodes where true values mean that at least one node can explain the positive sensor response. For a two-contamination, the following product can be done:

$$M = m_{ij} = \frac{\sum_{k=1}^{n} \beta(a_{ik}, a_{jk})}{n},$$

with n the number of rows of the simplified input/output matrix of contamination, the size of M is then potential node × potential node and

$$\beta(a, b) = (a \text{ OR } b),$$

a_{ij}being the coefficient of the simplified input/output matrix of contamination.

Each coefficient of the matrix M is the coupled result explaining positive responses at sensors. The coefficient represents the percentage of positive sensor responses that can be explained by the couple of nodes associated. The maximum coefficient of M is the one more likely associated with the couple responsible of those contaminations.

The diagonal coefficients m_{ii} of the matrix M are equal to the mean row values (see Table 8) when simplifying the input/output matrix of contamination. The Table 9 is the cross information matrix for a double contamination on nodes 5 and 15. It can be seen that their respective score 0.43 and 0.55 are not the highest on the diagonal (node 17 score is 0.74). A one contamination assumption would be, in this case, not placing the two contamination sources firsts.

Any coefficient m_{ij} not in the diagonal can't be inferior to either the two diagonal values m_{ii} and m_{jj} due to the OR operator maximization proprieties. If there is a large difference between diagonal elements and the maximum coefficient of M, the chance to have a multi-contamination is high. The highest value on the table is 0.96, which corresponds to the coupling of the two sources of contamination cited before. They can explain 96% of the positive sensor answers. And, as this value is quite above the two criteria values taken separately, it is most likely that this is a case of multi-contamination.

Table 9 : Multi-contamination cross information matrix.

Node Id	1	3	5	7	9	15	17	19	21	29	31	33	35	43
1	0.30	0.33	0.62	0.55	0.47	0.66	0.80	0.78	0.64	0.49	0.64	0.64	0.64	0.40
3	0.33	0.27	0.57	0.51	0.43	0.68	0.77	0.76	0.59	0.51	0.63	0.63	0.63	0.40
5	0.62	0.57	**0.43**	0.56	0.55	**0.96**	0.85	0.80	0.68	0.67	0.74	0.74	0.74	0.56
7	0.55	0.51	0.56	0.34	0.34	0.77	0.85	0.80	0.57	0.60	0.87	0.87	0.87	0.47
9	0.47	0.43	0.55	0.34	0.23	0.66	0.74	0.69	0.46	0.49	0.76	0.76	0.76	0.36
15	0.66	0.68	**0.96**	0.77	0.66	**0.55**	0.86	0.85	0.77	0.55	0.77	0.77	0.77	0.55
17	0.80	0.77	0.85	0.85	0.74	0.86	0.74	0.75	0.74	0.86	0.76	0.76	0.76	0.86
19	0.78	0.76	0.80	0.80	0.69	0.85	0.75	0.69	0.70	0.82	0.76	0.76	0.76	0.82
21	0.64	0.59	0.68	0.57	0.46	0.77	0.74	0.70	0.46	0.72	0.76	0.76	0.76	0.59
29	0.49	0.51	0.67	0.60	0.49	0.55	0.86	0.82	0.72	0.26	0.66	0.66	0.66	0.26
31	0.64	0.63	0.74	0.87	0.76	0.77	0.76	0.76	0.76	0.66	0.53	0.53	0.53	0.66
33	0.64	0.63	0.74	0.87	0.76	0.77	0.76	0.76	0.76	0.66	0.53	0.53	0.53	0.66
35	0.64	0.63	0.74	0.87	0.76	0.77	0.76	0.76	0.76	0.66	0.53	0.53	0.53	0.66
43	0.40	0.40	0.56	0.47	0.36	0.55	0.86	0.82	0.59	0.26	0.66	0.66	0.66	0.13

This method works for groups with more than two contaminations, however the time to calculate the coefficients is quadratic with the number of potential nodes, henceforth it is quite expensive for more than a simple couple. In other cases a coverage optimizing algorithm must be developed.

IV-1.3 Sensor Placement Criterion

The usual criteria for sensor placement are: average time to detection, detection likelihood, volume of water consumed, population exposed and extent of contamination. Some other criteria may be used such as detection redundancy but few in the literature take into account the performance of a method of source identification linked with a specific set of sensor locations.

The criterion for ranking explained in part IV-1.1 is modified and used in the following method to perform the sensor placement. The objective function, that is named Contribution, is defined as a real number between 0 and 1. It is defined as a non-dimensional ranking of the potential sources over the total number of nodes. It defines the capability of one sensor placement to rank the true source of contamination, as potential node, among the nodes list.

$$\text{Contribution} = 1 - \frac{\text{rank} - 1}{\text{node number} - 1}, \tag{34}$$

with node number being also the rank max.

If the contribution equals 1 then there are no nodes higher in the ranking, if it is 0 then the node is the least likely to be source of contamination. Also it has been chosen that when a node is not a potential source of contamination, it is given the Contribution 0.

Table 10: Contribution for a source identification with source of contamination being at node 3.

Node Id	3	5	7	9	17	19	21	23	31	33	35
Contribution	1.00	0.58	0.63	0.63	0.95	0.89	0.53	0.47	0.63	0.63	0.63

Table 10 gives the example associated with the results of Table 8. The node 3 is the first in the list, therefore it has the Contribution 1 associated with it. The others have an intermediate value between 0 and 1. In case of same criteria values, the smallest ranking is chosen, therefore nodes « 7 », « 9 », « 31 », « 33 » and « 35 » have equal rankings of 8. Then they are associated the value 0.63 equals to $1 - \frac{8-1}{20-1}$, 20 being the number of nodes in the network. The true source of contamination being the node 3, the criterion for the sensor placement is 1. If this would have been node 9 and the same values were found, the criterion would have been 0.63.

Therefore, from one sensor placement, and a given contamination, the method finds the Contribution of any potential source of contamination. The one corresponding to the true source of contamination can then be used to test the source identification method performance.

This criterion can also be used to perform the sensor placement method that is explained in the next section. The larger the Contribution, the higher is the source contamination potential node probability of being the true source of contamination. Depending on the contamination scenario, the sensor placement and the parameters BT and OT, three cases exist. The first one concerns the scenarios where the contamination is not detected, their Contribution is fixed to 0. The same applies for the scenarios where contamination source nodes are not inside the potential contamination nodes list. The accuracy of the source identification algorithm is then defined: the algorithm is accurate only if the source node is among the list of potential nodes of contamination:

$$\text{accuracy (\%)} = 100 \times \begin{cases} 0 \text{ if true source of contamination is in the potential contamination list} \\ 1 \text{ else.} \end{cases}$$

Lastly are the remaining contamination scenarios and the rank of the source of contamination among all the nodes. The specificity of the source identification algorithm is then defined as follows:

$$\text{specificity (\%)} = \left(1 - \frac{\text{average ranking} - 1}{\text{node number} - 1}\right) \times 100$$

and represents the percentage of nodes that are ranked worse than the source of contamination. A 100% specificity indicates that the source node is ranked first.

The scenarios of contamination belong either in the not detected scenarios, the not accurate or the ranked scenarios.

$$N_{scenarios} = N_{not\ detected} + N_{not\ accurate} + N_{ranked}$$

The average Contribution can be calculated with the following formula:

$$\overline{\text{Contribution}} = \frac{\sum_{i=1}^{N_{ranked}} \left(1 - \frac{\text{rank(i)} - 1}{\text{node number} - 1}\right)}{N_{scenarios}}.$$

It can also be calculated as a function of the specificity percentage, the accurate percentage and the detection likelihood, with the following formula:

$$\overline{\text{Contribution}} = \frac{\text{specificity percentage}}{100} \times \frac{\text{accurate percentage}}{100} \times \frac{\text{detection likelihood}}{100}.$$

IV-2 Source Identification Results

A sufficiently large number of simulations of random one-contamination have been performed for the test network which has around 2,500 nodes. Each simulation lasts for 72 hours, and the contaminations are simulated between 24h and 48h for durations of 1h to 6h. Simulations with a fixed concentration of 1mg/L of conservative contaminant have been taken to simulate sensor responses. The sensors are considered respecting the following hypothesis: any contamination value that is higher than 0.001 mg/L in the simulation is taken as a positive response, otherwise it is negative. The time of backtracking BT is taken as 24h and the Contribution criterion will be plotted as a function of the time of observation OT for both average time detection optimal placement (ATDOP) and detection likelihood optimal placement (DLOP) given by Piller et al. (2015).

IV-2.1 Ranking for 10 Sensors

First 10 sensors have been used. In Figure 52 Contribution for both average time detection (ATDOP) and detection likelihood (DLOP) optimal placement with 10 sensors has been plotted as a function of OT. For both, the Contribution increases with OT, proving the efficiency of exploiting new information by the source identification algorithm. The DLOP sensor placement is doing better than ATDOP. Indeed it has a higher probability of detection and therefore is less impacted by the constraints. The Contribution is equal to 0 if not detected.

When looking at the accuracy and the specificity, ATDOP is doing slightly better. In Figure 53, the accuracy is increasing from 95 or 96% to 99% with OT increasing. Most contamination scenarios that are not backtracked, and therefore not accurate, are probably because that the contamination happened before the backtracked time. This explains why ATDOP, which is the average time to detection optimal placement, performs better than the detection likelihood optimal placement when observation time is low. Table 11 5 shows, as expected, that ATOP has an average time to detection lower than DLOP but also has a lower detection likelihood. In Figure 53, the specificity is going from near 93% to at most 96%, which is correct, the source node is on average among the first 7% nodes which is equal to at most 175 nodes.

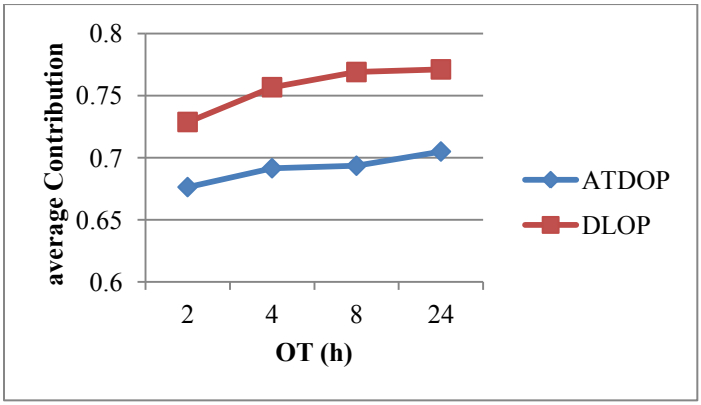


Figure 52: Contribution for average time to detection and detection likelihood optimal sensor placement with 10 sensors.

Table 11: Average time to detection and detection likelihood scores for ATDOP and DLOP.

	ATDOP	DLOP
Average time to detection	3.8h	6h
Detection likelihood	75%	82%

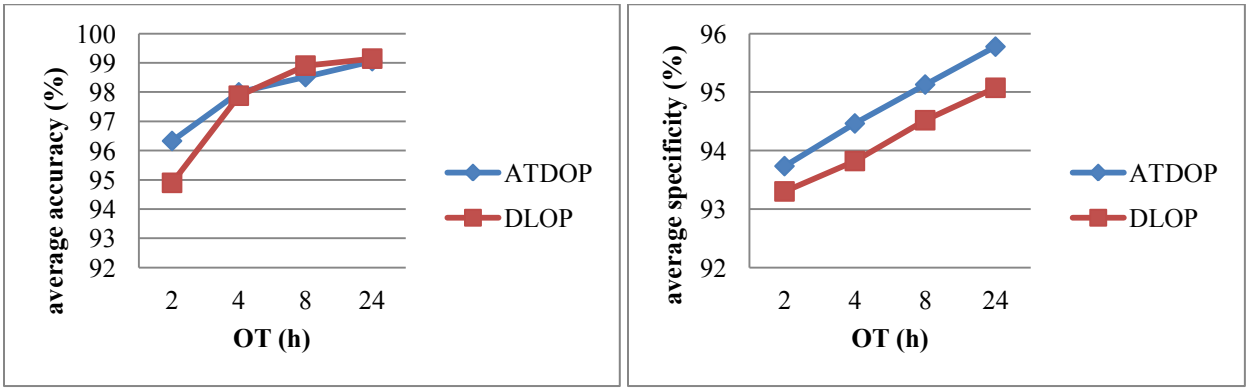


Figure 53: Accuracy and specificity for average time to detection and detection likelihood optimal sensor placement with 10 sensors.

IV-2.2 Ranking for 10, 20 and 50 Sensors

It is also interesting to see the influence of the number of sensors on the ranking. In Figure 54 the Contribution for both average time to detection and detection likelihood optimal placement have been plotted as a function of sensor numbers. The Contribution increases with the number of sensors for both sensor placements. This increase is not linear, the gain from 10 sensors to 20 sensors is not the same as between 20 and 50 and the asymptotic value seems lower for average time to detection optimal placement than for detection likelihood optimal placement.

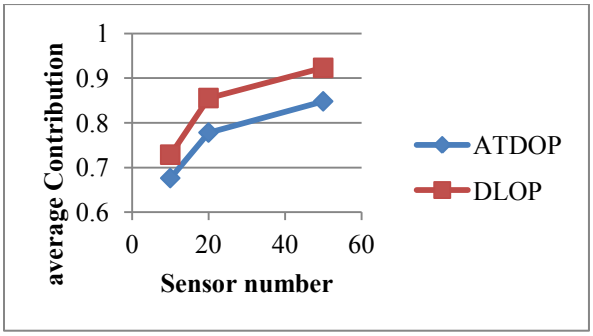


Figure 54: Contribution for average time to detection and detection likelihood optimal sensor placement with 10, 20 and 50 sensors.

In Figure 55, the Contribution is also given but this time as a function of the observation time for 10/20/50 sensors. For average time to detection optimal placement, the more sensors there are, the less is the result is impacted by the observation time. A similar behavior can be seen with the detection likelihood optimal placement but with less sensitivity. Finally, average time to detection optimal placements are doing better than respective detection likelihood optimal placements for both accuracy and specificity as seen in Figure 56. The sensor number impacts greatly on the source identification results but with less impact for each new sensor added, and the decision to choose the number of sensors to place depends on the four desired criteria: the type of optimization, the Contribution, the average time to detection and the detection likelihood, as given in Table 12.

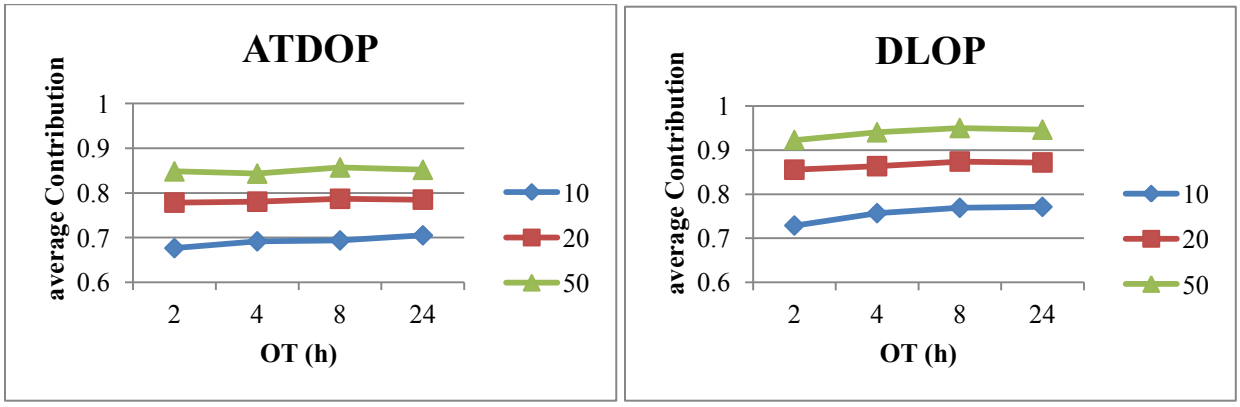


Figure 55: Contribution for average time to detection and detection likelihood optimal sensor placement with 10/20 and 50 sensors in function of observation time.

Table 12: Average time to detection and detection likelihood for ATDOP 10/20/50, DLOP 10/20/50.

	ATDOP 10	ATDOP 20	ATDOP 50	DLOP 10	DLOP 20	DLOP 50
Average time to detection	3.8h	2.7h	1.8h	6h	4.4h	3.4h
Detection likelihood	74%	80%	84%	82%	91%	95%

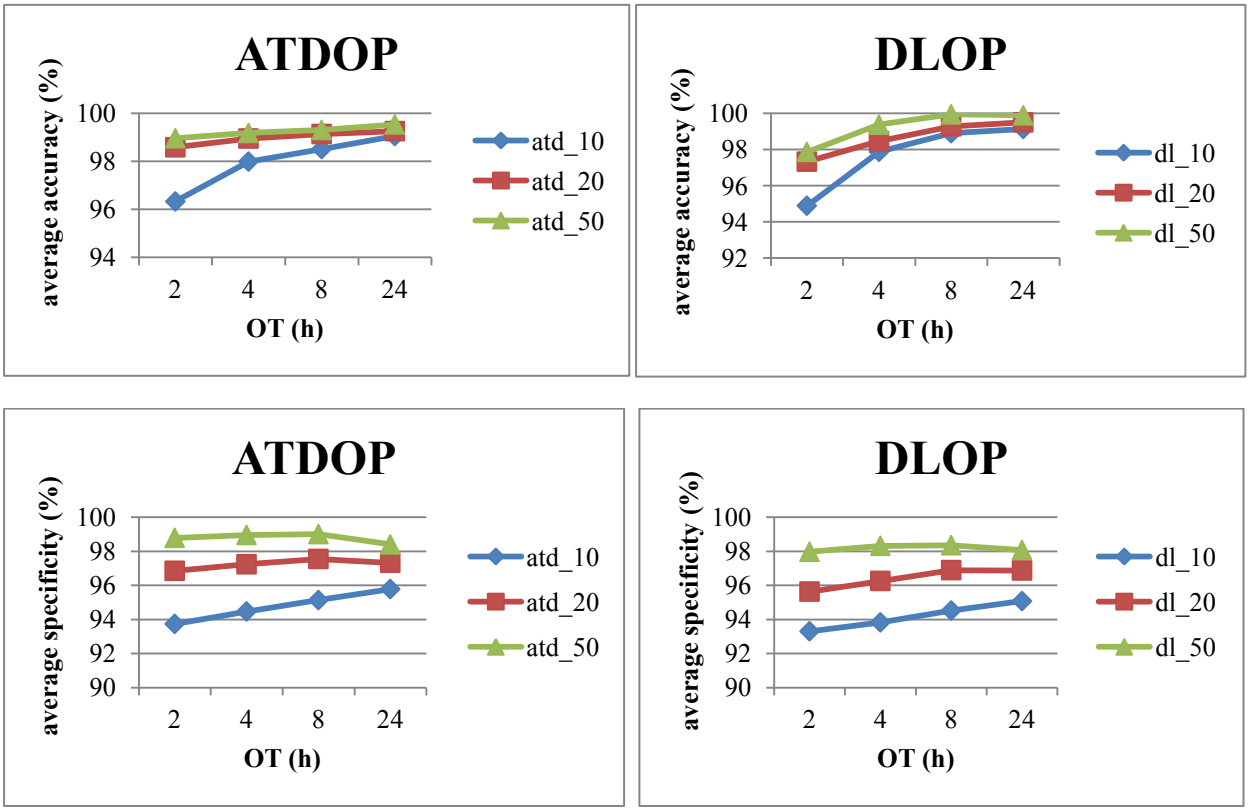


Figure 56: Accuracy and specificity for average time to detection and detection likelihood optimal sensor placement with 10, 20 and 50 sensor.

IV-3 Conclusion

The novel method of source identification uses a backtracking algorithm to construct the input/output contamination matrix in real time. The backtracking is based on the adjoint of the transport equations. It computes those in reverse time, beginning at sensors, to enumerate potential nodes of contamination. The use of the backtracking methods allows to handle large-size networks that would require huge calculation times if it was performed with a forward scheme. Indeed, the time needed to run every possible contamination is exponential with the size of the network. The matrix obtained is then analyzed (with simple manipulation such as aggregation) to evaluate the number of positive sensor responses that each potential source of contamination can explain. This score allows the creation of a ranking list among the potential sources of contamination. The method uses positive answers to find potential nodes but it also processes negative answers to eliminate candidates. It can also be extended to multi-contamination but is quite costly in time; extended covering methods may be developed to overcome this drawback.

The method has been optimized to be computed in real-time condition even for big network and a large number of sensors. The backtracking algorithm contains break conditions in case of repetitions: graph edges are not checked twice for same time and sensor response. In result, the method takes at most several minutes to compute even for a big network and a widespread contamination.

To take into consideration detection likelihood, the accuracy and the specificity criteria in the sensor placement, one new criterion is defined that is called Contribution. This is a dimensionless variable between 0 and 1 that depends on the ranking, the last ranked are given the value 0 and the first to be source of contamination is given the value 1. That criterion is averaged on several contamination scenarios. The not detected scenarios as well as the not accurate scenarios are given the value 0. The Contribution is then the average of the results given by all the contamination scenarios simulated.

The methods have been tested on a real French network with around 3,000 pipes. The influence of the backtracking time (BT) as well as the initial sensor placement for the local search method has been carried out. Eight parameter/method combinations have been defined: Two come from the application of the Piller et al. (2015) method each based on average time to detection (ATDOP) and detection likelihood (DLOP) maximization through a greedy algorithm. Six method/parameter scenarios have been added to ATDOP and DLOP. The following two come from a greedy algorithm but with objective being the maximization of the Contribution criteria in the two parameter cases BT being long, 24h, or short, 4h, and observation time (OT) being short set at 2h. The fourth last sensor placements assessed are results of a local search

algorithm in the same configuration than for the greedy algorithm but with two different initializations. The first twos begin with ATDOP, the other two launched from DLOP.

It was found that the criteria are conflicting. When comparing ATDOP and DLDOP, the first one is performing better with average time to detection, accuracy and specificity, however it performs worse concerning detection likelihood than the second one. It can be concluded that ATDOP performs better the source identification but on less coverage of the network.

Also, adding new sensors to ATDOP and DLDOP improves the results for the criteria, however each new added sensor brings fewer result improvements each time. Particularly ATDOP gets less detection likelihood improvement than the other placement, impacting the overall score for the source identification performance.

In this chapter we have defined four variables upon which the source identification gives different results depending on the sensor placement it is applied on: average time to detection, detection likelihood, the accuracy and the specificity. The first twos are conflicting and therefore an optimization problem solving is needed. The average time to detection will be constrained through the BT, the others criteria through OT and the objective function being the minimization of the Contribution.

CHAPTER V

Sensor Placement Adapted to Source Identification

The aim is to look for the best placement of sensors in a water distribution network to monitor and detect contaminations. In this chapter we will consider the solving of an optimal optimization problem with averaged Contribution as objective function.

V-1 Models

The optimization problem is solved by two methods. The first method uses a coupling between the backtracking method presented before and a Monte Carlo method (for contamination scenarios) associated with a greedy algorithm. Then a local search on graph method is also given which is a quicker algorithm.

The results are then analyzed and compared between them as well as the solutions given by Piller et al. (2015). It is a greedy algorithm that takes into account several computed quality parameters to define the best location for sensors placements. Among those are the average time to detection and the detection likelihood.

V-1.1 Greedy Algorithm

Piller et al. (2015) perform a several mono-objective optimization solving on conflicting criteria such as minimization of time to detection, maximization of the detection likelihood and minimization of the fraction of population exposed. The method uses a greedy algorithm which is easy to compute. However, in some cases, it may not give the optimal solution. It has to be noted however that Propato et al. (2006) showed near optimality for this greedy algorithm. The aim of the optimization is to place the sensor so as to respect as much as possible the criterion chosen. A criterion defined as a weighted sum of the simple criteria can also be used instead.

It is proposed to add a new criterion, which optimizes the location that performs the best for source identification. It is defined as the placement that gives the best ranking to the real source of contamination-associated candidates. It uses the criterion defined in part IV-1.3, based on the ranking among all nodes of the network, that is the Contribution of one sensor placement for the localization of the source of contamination. The method follows the flowchart in the Figure 57. Firstly several contamination scenarios are defined as well as an empty list of sensors. All nodes are added to this list separately as temporary sensors and all contamination scenarios are tested giving the average Contribution for that temporary sensor. The node with the highest Contribution is then added permanently to the list of sensors. Then the greedy algorithm can be used again, until the number of sensors corresponds to the value wanted.

An example of the result is given in Table 13, only the first 10 nodes and 11 (among 20) contamination scenarios are shown. Each cell gives the value of the Contribution of the potential

sensor associated to the contamination scenarios. Here, node 47 on average the one which gives the best results. Then after several use of the algorithm, node 35, and then 49 are chosen. These are in the same order in Table 13 but may not be for other examples.

Table 13: Average Contribution calculations for 20 scenarios.

	NodeId	47	35	49	33	45	31	23	19	7	21
Scenario	Contamination node										
1	19	0.53	0.53	0.42	0.58	0	0	0.63	0.74	0.63	0.68
2	23	0	0	0	0	0	0	0.74	0	0	0
3	51	0	0	0	0	0	0	0	0	0	0
4	47	0.95	0	0.83	0	0	0	0	0	0	0
5	23	0	0	0	0	0	0	0.63	0	0	0
6	3	0	0	0	0	0.63	0.74	0	0.74	0.58	0.68
7	19	0.84	0.79	0.67	0.84	0	0	0.89	0.95	0.74	0.89
8	9	0	0	0	0	0	0	0	0	0.58	0
...	...	...	...	...	...	...	...	...	...	...	...
18	51	0	0	0	0	0	0	0	0	0	0
19	31	0.53	0.53	0.42	0.58	0.63	0.73	0	0	0	0
20	31	0.53	0.53	0.42	0.58	0.63	0.73	0	0	0	0
	mean	0.28	0.27	0.27	0.25	0.23	0.22	0.22	0.20	0.19	0.19

The algorithm for sensor placement is presented in Figure 57 and is the following:

- begin with a list with no sensor in the network

- calculate for every node its average score for the criterion Contribution by a Monte Carlo method when adding temporarily the node to the list of sensor.

- add the best sensor to the list of sensors

- keep on until the number of sensors is reached

It is an effective method, however it is quite expensive. The method has only been used for cases of one contamination. For multi-contamination, more computation time is needed and the algorithm has to be modified.

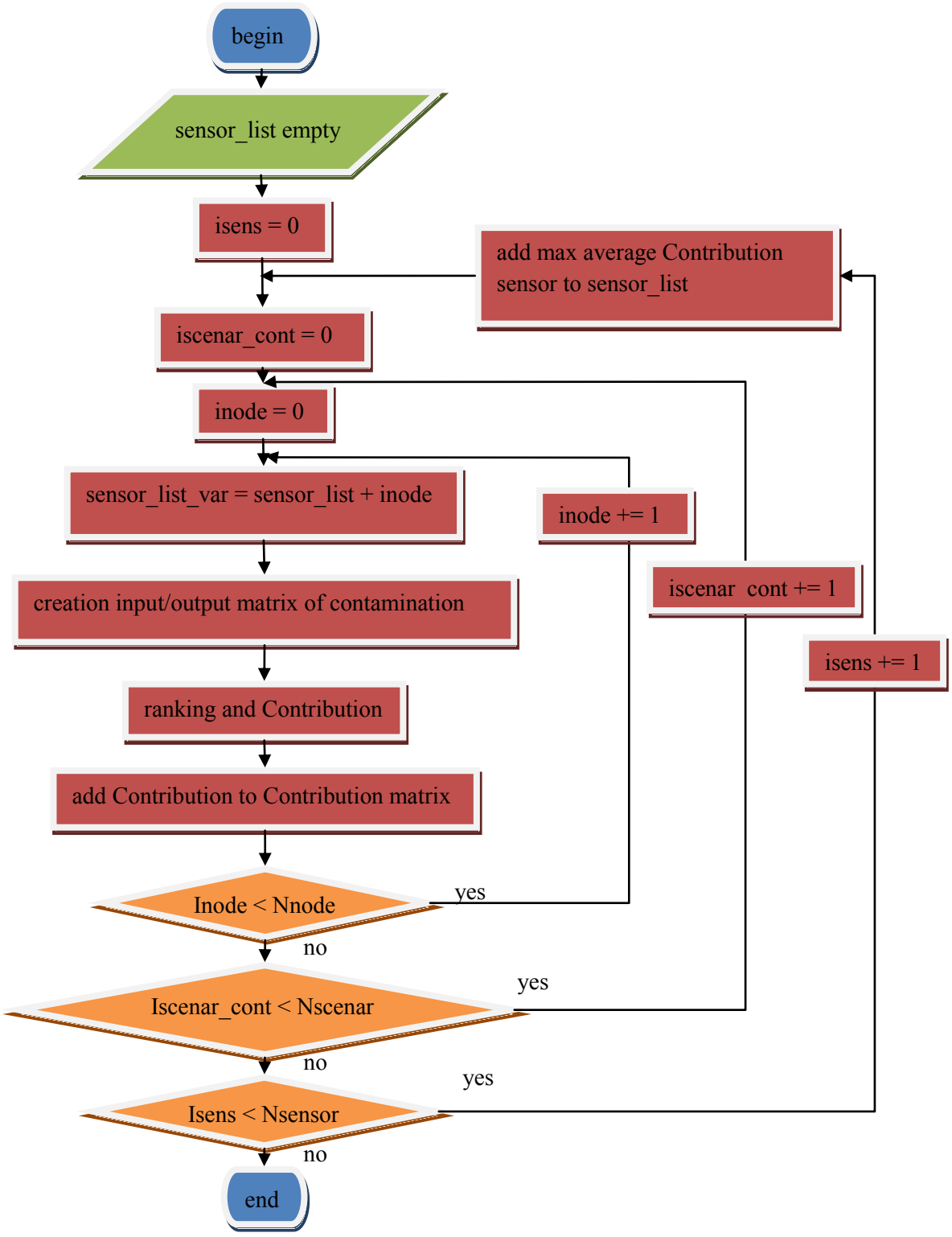


Figure 57: Algorithm of sensor placement.

V-1.2 Local Search on Graph Algorithm

The previous algorithm looks at each iteration for the node that best suits the criteria of source identification. As the search is done on the whole node list, this method is quite costly as shown in the next part. To accelerate the algorithm, an optimization based on local search on graphs method is used. It begins with an initial set of guessed best locations, which can be a list given by another method such as the average time detection optimal placement given by Piller et al. (2015). Then the optimal solution is found by testing the Contribution source identification criteria on adjacent nodes. If the criterion has increased on adjacent nodes compared to the actual node then it is taken as the new sensor location in the next iteration. It may not converge to one solution (as seen later in Figure 61 and Figure 62) because all potential sensors are changed at the same time for each iteration to accelerate the algorithm. A convergence criterion can be defined to stop the algorithm and returns the best placement found. It is also proven later that it may not give the optimal solution but local minima. To improve the method, adjacent to already-used adjacent nodes can be explored or a genetic algorithm be used. In this thesis we have taken a three adjacent nodes distance.

V-1.3 Parallelization

Parallelization of the code has been necessary because of the large number of contamination scenarios that have been taking too much computational time. Both precedent algorithms have been parallelized and launched on a cluster for computation. For the first algorithm, for each potential sensor node 2,000 contamination scenarios are tested on a test network case. Firstly these 2,000 cases are divided into 20 groups of 100 because of the memory limit of the processors needed for large networks. The contamination scenario groups are then divided by groups of processors depending on their numbers. Each processor is given a certain amount of scenarios to process. These scenarios are taken from the 100 scenarios known by that processor. To accelerate the algorithm an optimization has been performed on the work list of each processor. To each scenario, a certain amount of time is associated that depends mostly on the scenario and on the sensor placement, therefore a translation on the scenario list has been applied, and the same scenario might not be done by the same processor for another potential sensor. This allows us to average the calculation time charge for all processors, decreasing the total time of calculation. Indeed, for each iteration, all processors need to synchronize to decide on the sensor to add to the list, therefore the total time of calculation is based on the processor that takes the most time to calculate, and that mostly depends on the contamination scenario it needs to perform.

V.2. Sensor Placement Results

The Contribution criterion studied in the previous section is now used as an objective for the optimal sensor placement problem. The greedy and local search on graph methods have been used on the test network. Two cases have been taken, BT = 4h, OT = 2h and BT = 24h, OT = 2h with 10 sensors. The results are presented hereafter.

V.2.1 Greedy Algorithm Results

The test network has been tested for the greedy algorithm. Two cases have been performed BT = 4h, OT = 2h (GOP4) and BT = 24h, OT = 2h (GOP24) with 10 sensors. It consists of testing every potential sensor location and determining, for all generated contamination scenarios, which location is best for the placement of the next sensor. Concerning the Contribution (see Figure 58), both DLOP and GOP24 are performing best compared to GOP4 in third and ATDOP in last place. The Table 14 gives the average time to detection and the detection likelihood of the two sensor placements. The detection likelihood is in the following ascending order ATDOP, GOP4, GOP24 and DLOP. Concerning the accuracy and specificity (see Figure 59), GOP4 performs better and equal respectively with GOP24. Depending on the wanted criteria, either average time to detection, detection likelihood, accuracy or specificity will be chosen GOP4, DLOP, GOP4 or GOP24 respectively.

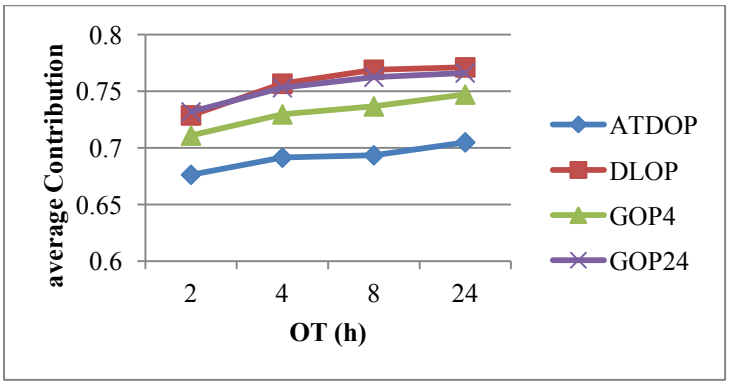


Figure 58: Contribution for average time to detection, detection likelihood and greedy algorithm optimal sensor placement with 10 sensors.

Table 14: Average time to detection and detection likelihood for ATDOP, DLOP, DOP4 and GOP24.

	ATDOP	DLOP	GOP4	GOP24
Average time to detection	3.8h	6h	3.5h	5.1h
Detection likelihood	74%	82%	76%	80%

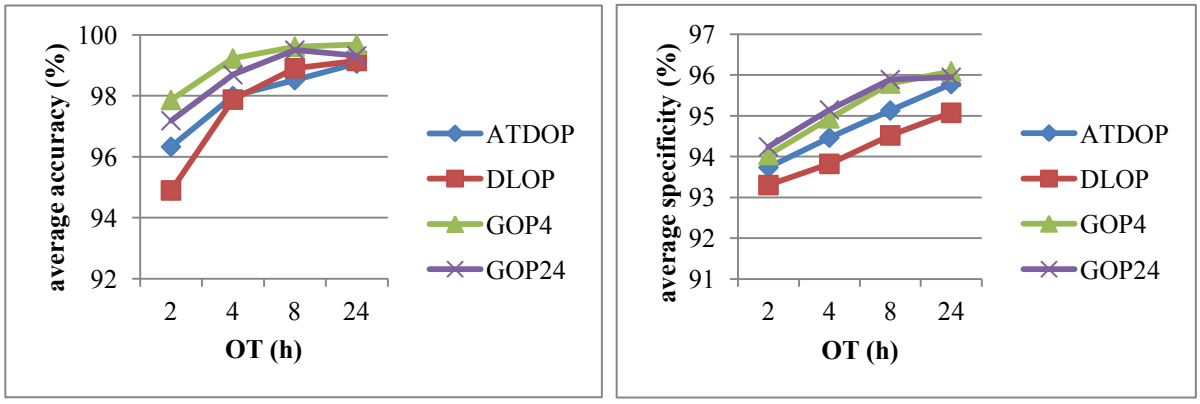


Figure 59: Accuracy and specificity for average time to detection, detection likelihood and greedy algorithm optimal sensor placement with 10 sensors.

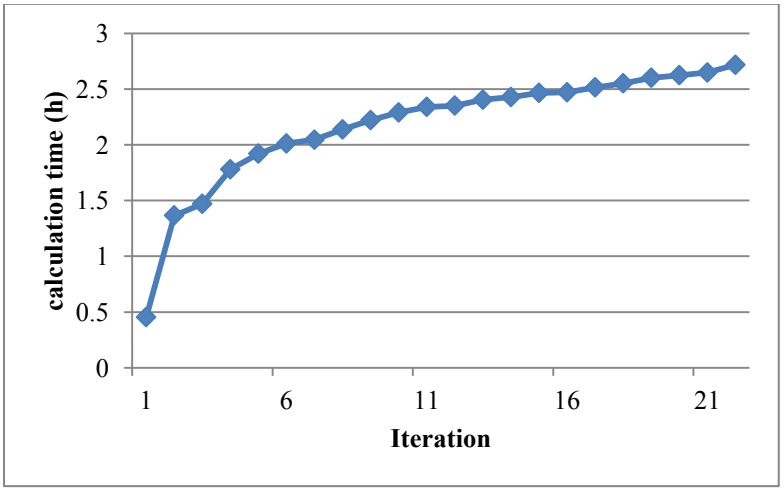


Figure 60: Execution time per iteration for the greedy algorithm with BT = 24h and OT = 2h with 120 processors for 23 sensors.

Figure 60 shows the time it takes, at each iteration, to find a new sensor location with 120 processors, which is not the cumulated time. The first iteration takes less than half an hour, then the execution time increases, until it reaches an asymptotic slope, adding less than 10 minutes for each new sensor from the last execution time. The time of execution increases with the number of sensors set because the complexity increases. As sensor placement is not done in real time conditions, the execution times are still not a problem on a sufficient cluster. However it is believed that for bigger networks it might be problematic because the problem is NP hard. Another algorithm, based on local search on graph is developed to decrease the execution time.

V-2.2 Local Search Algorithm

A local search algorithm with 3 adjacent node distance has been used for 10 sensors. At each iteration, a local search around the 10 potential sensors location is performed concerning the source identification criteria. Each potential sensor is replaced with the local neighbor that performed the best. Firstly the evolution of the criteria is given, defined in the first section, as a function of the iteration. Then the last sensor placement performance is tested for source identification for both initialization with average time to detection and detection likelihood optimal placement.

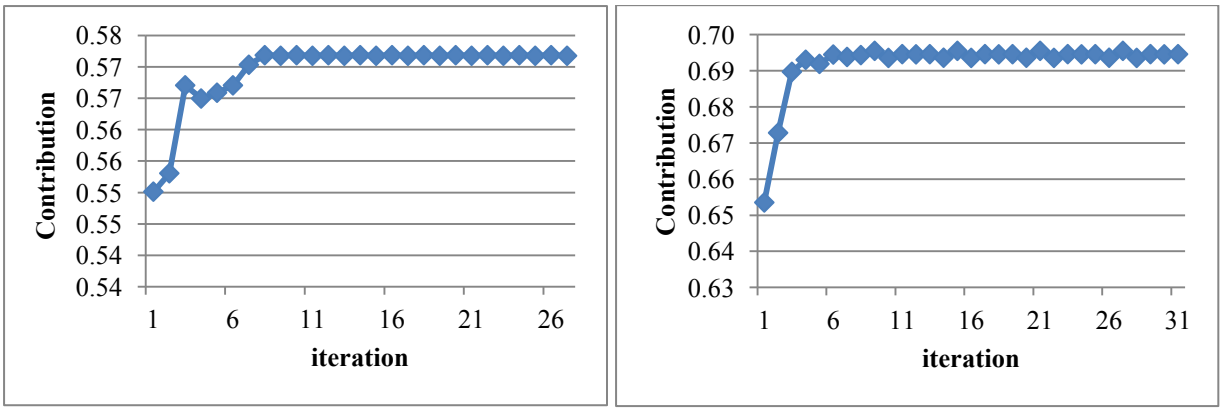


Figure 61: Criteria evolution in function of iteration for local search with average time detection initialization and BT = 4h or BT = 24h and OT = 2h for 10 sensors.

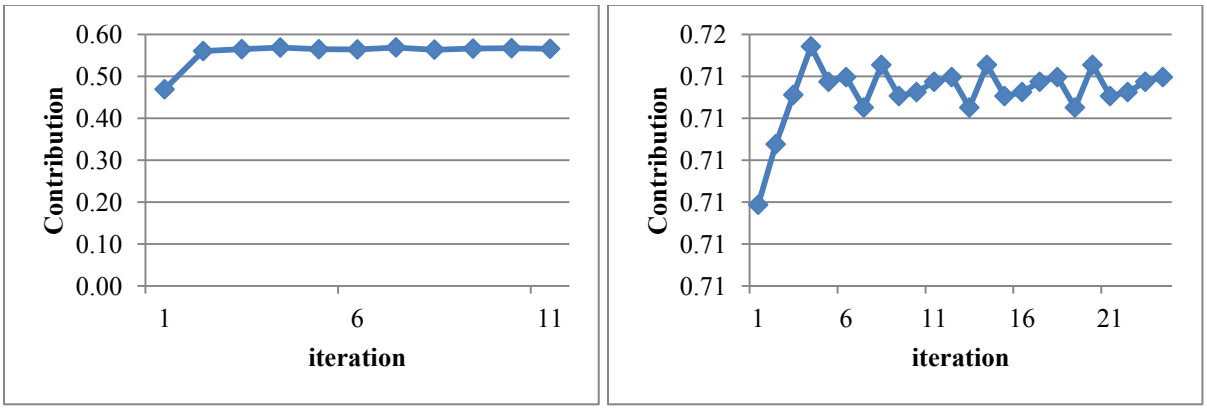


Figure 62: Criteria evolution in function of iteration for local search with detection likelihood initialization and BT = 4h or BT = 24h and OT = 2h for 10 sensors.

Four cases have been performed BT = 4h, OT = 2h (LSATDOP4, resp. LSDLOP4) and BT = 24h, OT = 2h (LSATDOP 24, resp. LSDLOP24) with ATDOP (resp. DLOP) initialization and 10 sensors.

In Figure 61 and Figure 62 the evolution of the Contribution is shown, defined in the first section, as a function of the iterations. It can be seen that the algorithm has converged in all cases in fewer than 10 iterations. The Contribution values are slightly better when the initialization is taken as detection likelihood optimal sensor placement. The time for each iteration is constant around 700 seconds for 60 processors and therefore the algorithm converges in more or less one hour, which is a lot faster than the previous algorithm.

The Contribution scores in ascending order are ATDOP, LSATDOP4, LSATDOP24, LSDLOP4, LSDLOP24 and DLOP. The Contributions values are between the average time to detection and detection likelihood optimal placement Contributions. The same applies for average time to detection and detection likelihood. Finally, LSATDOP24 and LSATDOP4 perform best concerning accuracy and specificity. As discussed before, the Contribution criteria is not enough to decide which sensor placement to choose, detection likelihood, average time to detection, accuracy and specificity are also decision factors. If average time to detection is favored, LSATDOP4 may be chosen because it has better results than ATDOP and almost the same average time to detection. In case of favouring detection likelihood, LSATDOP24 and LSDLOP24 may be choosen because they have better accuracy and specificity than DLOP. LSDLOP4 is good on average for average time to detection and detection likelihood. In conclusion this method is giving better solutions overall than initial ones and which parameters and initialization to use depend on the results we want to favor.

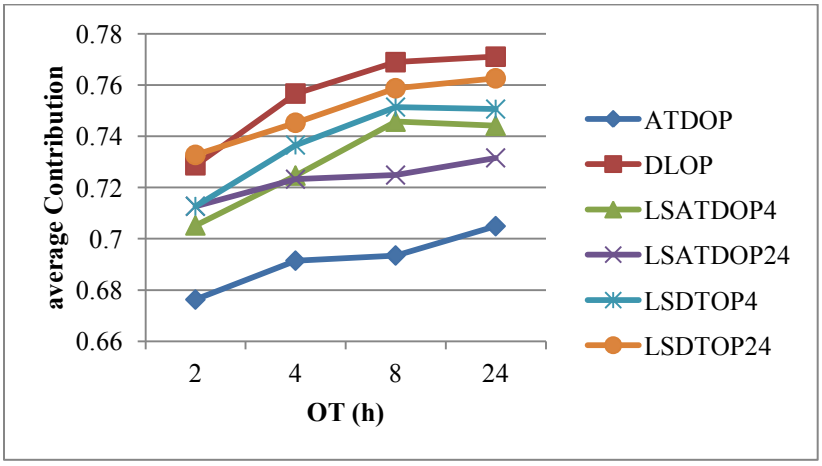


Figure 63: Contribution for average time to detection, detection likelihood and local search algorithm optimal sensor placement with 10 sensors.

Table 15: Average time to detection and detection likelihood for ATDOP, DLOP, LSATDOP4, LSATDOP24, LSDLOP4 and LSDLOP24.

	ATDOP	DLOP	LSATDOP4	LSATDOP24	LSDLOP4	LSDLOP24
Average time to detection	3.8h	6h	3.9h	5.4h	4.3h	5.2h
Detection likelihood	74%	82%	77%	79%	80%	81%

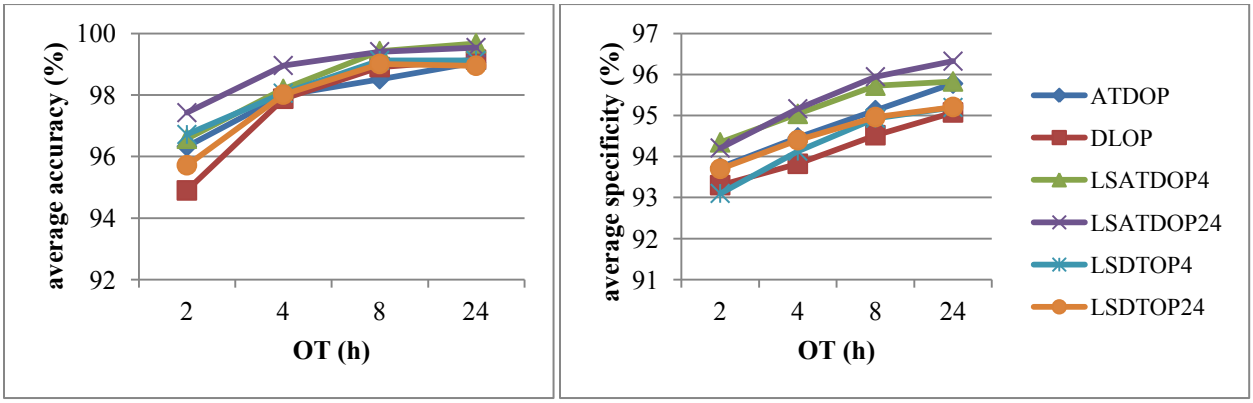


Figure 64: Accuracy and specificity for average time to detection, detection likelihood and local search algorithm optimal sensor placement with 10 sensors.

V.2.3 Evaluation

We have seen that different objectives can influence the choice for a source identification optimal sensor placement: the average time to detection, the detection likelihood, the accuracy, the specificity and finally the Contribution which is a combination of the last three. The rankings for all these parameters have been written in Table 16 and the mean over all parameters is given with BT = 24h and OT = 2h. Overall, the greedy algorithm solutions are the best followed by the local search algorithm solution LSATDOP4, LSATDOP24 and LSDLOP24. This ranking should however be used cautiously and different weights can be used depending on which is considered more important for the decision maker.

Table 16: Evaluation for all objectives.

	Average time to detection rank	Detection likelihood rank	Accuracy rank	Specificity rank	Contribution rank	Mean rank
GOP24	5	3	3	2	2	**3**
GOP4	**1**	7	**1**	4	6	3.8
LSATDOP24	7	5	2	3	5	4.4
LSATDOP4	3	6	5	**1**	7	4.4
LSDLOP24	6	2	7	6	**1**	4.4
LSDLOP4	4	4	4	8	4	4.8
DLOP	8	**1**	8	7	3	5.4
ATDOP	2	8	6	5	8	5.8

To conclude this section, we have described two methods: a greedy algorithm and a local search on graph to maximize the Contribution criterion. Both were better for source identification (accuracy and specificity) than optimization based on average time to detection and detection likelihood. However these methods are time-consuming. The greedy algorithm execution time is between 30 minutes to 3 hours on 120 processors with each new sensor added. The local search on graph showed best results with both average time to detection initialization, BT = 4h, BT =

24h, OT = 2h criteria and detection likelihood initialization, BT = 24h, OT = 2h criteria. The algorithm converges in approximately one hour in total on 60 processors. The local search solutions are not as good as the greedy ones but can be chosen as their results are not too far and the time of execution is a lot faster.

V.3 Conclusion

For a given sensor placement, the forward transport method is used to evaluate two criteria of the battle of the water sensor networks BWSN which are the average time to detection (ATD) and the detection likelihood (DL), chosen because they are worldly used nowadays in early warning sensor placement. Additionally, the backtracking method is used and two other criteria have been defined that evaluate the source identification Contribution. Among the scenarios of pollution simulated, the contamination is detected but the source is computed as not being in the list of potential source of contamination, because of insufficient backtracking time and numerical inaccuracies; the percentage of those scenarios among the detected scenarios is given as the accuracy percentage. Finally, a last criterion is taken which is the specificity which is calculated from the rank of the true source of contamination among the scenarios remaining.

The Contribution criterion can then be used as an objective function to be maximized and two method solutions of sensor placements were tested. Both methods use a Monte Carlo algorithm to simulate contamination events inside a network and measure their impacts. The first one uses a greedy algorithm and tests every node as potential sensor choosing at each iteration the best candidate which is added to the list of sensors and reiterates until the amount of sensors desired is obtained. The second method uses a local search on graph algorithm testing three adjacent distance nodes on the Contribution criterion for source identification. It iterates until the maximum Contribution is found.

Six method/parameter configurations have been added to ATDOP and DLOP. The following two come from a greedy algorithm but with objective being the maximization of the Contribution criteria in the two parameter cases BT being big, 24h, or little, 4h, and OT being short set at 2h. The last fourth sensor placements assessed are results of the local search algorithm in the same configuration than for the greedy algorithm but with two different initializations. The first two ones begin with ATDOP, the other two ones launched from DLOP.

Both methods have been parallelized and launched on a calculation cluster. The two methods use two parameters that influence the result of the source identification. The first one is the time of backtracking, how much time the algorithm search in the past from the first time to detection (BT). The second is the time of observation after the first time of detection (OT) it gives the source identification algorithm to process information.

The optimal sensor placement methods adapted to the source identification Contribution criteria are doing better with the new source identification overall, than the first two sensor

placements, ATDOP and DLOP, however they take more time to execute. Also, the greedy method is doing better than the local search method but is slower.

Concerning the backtracking time BT parameter, a small value (e.g. 4h) induces a small average time to detection score. This parameter is important, because the results show that a small average time to detection is performing better but on a smaller scale than with a big detection likelihood. For the observation time OT, 2h seems to be enough to have an idea to identify the contamination, still a bigger observation time improve the results.

In conclusion of the two methods of sensor placement, greedy and local search: the placements based on the first method presented in this chapter, greedy algorithm with Contribution objective, are doing the best overall. However they are the longest with an increasing iteration execution time going from half an hour to three hours, using one hundred twenty processors. The solutions given by the local search are second and the method is much faster than the previous one taking three hours overall on sixty processors. Finally ATDOP and DLDOP are less suited for source identification than the two previous methods but are fast to compute.

The choice of the method will depend on the objectives and the network properties as well as the number of sensors available. More research work should focus on evaluating the impact of hydraulic and transport model error in the final optimal designs. Firstly, the velocity may not be accurate or not the same as in the simulations used to place the sensors. Secondly, transport model, reaction coefficients and sensor thresholds need careful attention. Results from the SMaRT-OnlineWDN project have shown that imperfect mixing at cross junctions and double T-junctions may have an impact on the contamination spreading. The same applies when adding the dispersion effect to the transport model. The two modifications can influence the simulation of the contamination scenarios, therefore it can also change the sensor responses, and finally it influences the sensor placement optimization.

Also, for exploring the merits of sensor optimal placements, more statistics, other than the mean, can be conducted on the results of the different Monte Carlo scenarios to workout robust dispersion indicators such as the interquartile range. It will permit to make clustering among equivalent sensors solutions.

CHAPTER VI

Contamination Case Study

The sensor placement adapted to source identification has been used on a case study network. It is a large network composed of more than 13,000 nodes and four district metering areas (DMA). Because of the size of the network, all methods developed previously will take a lot of time. Also the network being bigger means that more sensors are needed and also more sensor responses are to be proceeded for the source identification algorithm. In this case, the number of sensor to look for is set as 50. That represents one sensor every 20 km of pipes, which is a good value for current monitoring of networks. Also for security measures, networks will not be shown and will be replaced with blank color filled filtered images.

VI-1 Sensor Placement and Source Identification Efficiency for the Large Case Study

First the greedy algorithm, presented in V1.1, cannot be used as it is too slow, not even one sensor is calculated with 1,200 processors during 4 hours. The average time to detection and detection likelihood optimal placement will be compared to the result given by the local search on graph algorithm, which has been computed on 60 processors for a day.

The Figure 65 shows the Contribution results for the different sensor placements in function of the time of observation. It was shown in the previous part that the Contribution criterion is strongly connected to detection likelihood criteria. Indeed, scenarios where the contamination is not detected are given the Contribution zero. To evaluate a sensor placement, four criteria are defined in previous chapter: average time to detection, detection likelihood, not accuracy and specificity. The equation (35) gives the relations between those criteria and the Contribution.

ATDOP and DLOP have been obtained with Piller et al. (2015) for a first set of contaminations: a duration of 120h with contamination starting the first day. However, those sensor placement are assessed (average time to detection, detection likelihood and the other results) from scenarios of contamination with a simulation duration of 72h with the injection the second day. Therefore, ATDOP and DLOP are not optimal for those new conditions. Indeed, Table 17 gives the average time to detection and detection likelihood of the 6 sensor placements evaluated in this chapter, and DLOP has the lowest detection likelihood.

The choice of the new set of contamination is motivated by two reasons. The first one is because a smaller time of simulation will induce smaller time of execution from the sensor placement algorithms. The second one is because we want to be close to the reality. 24h is already a lot of time for a contamination to spend in the network. In this network, the maximum residence time is close to 5 days, therefore a good detection likelihood optimal sensor placement

with no restriction will tend to be at the borders of the network. However, such sensor placement will have a poor detection rate for an observation time like 24h.

In Table 17, the ATDOP and the local search that takes it as initialization are doing better with a detection likelihood around 70% and average time to detection around 4h for ATDOP and LSATDOP4, and 5.6h for LSATDOP24. The local search placements with DLOP as initialization have a better Contribution with a detection likelihood around 78% and average time to detection 5.5h and 6.2h. However, by comparing the different sensor placements on accuracy and specificity (see Figure 66), the local search based on ATDOP and backtracking time of 4h is doing the best for the two criteria. And by evaluating all sensor placements on all criteria, LSATDOP4 is the one that gives the best score overall as shown in Table 18.

Therefore, the LSATDOP4 will be the one sensor placement that will be used in the following.

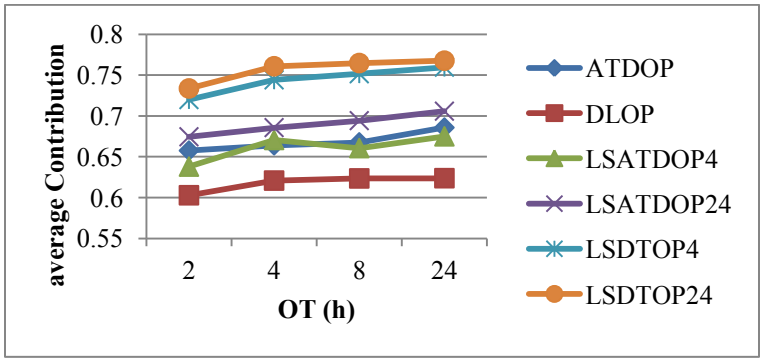


Figure 65: Sensor placement Contribution comparison on test case network.

Table 17 : Average time to detection and detection likelihood for ATDOP, DLOP, LSATDOP4, LSATDOP24, LSDLOP4 and LSDLOP24 on Case study network.

	ATDOP	DLOP	LSATDOP4	LSATDOP24	LSDLOP4	LSDLOP24
Average time to detection	4.3h	4.3h	4.1h	5.6h	5.5h	6.2h
Detection likelihood	70%	64%	69%	72%	77%	79%

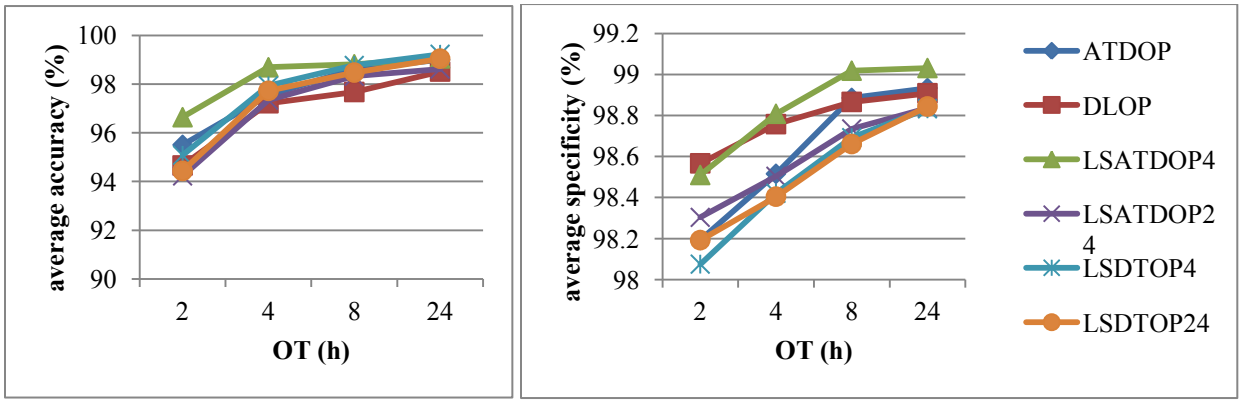


Figure 66: Accuracy and specificity for greedy and local search optimal sensor placement with 10 sensors.

Table 18 Evaluation for all parameters on case study network for OT = 2h.

	Average time to detection rank	Detection likelihood rank	Accuracy rank	Specificity rank	Contribution rank	Mean rank
LSATDOP4	1	5	1	2	5	2.8
ATDOP	2	4	2	4	4	3.2
LSDLOP4	4	2	3	6	2	3.4
LSDLOP24	6	1	5	5	1	3.6
LSATDOP24	5	3	6	3	3	4
DLOP	3	6	4	1	6	4

ATDOP is second in the evaluation, still LSATOP4 gives better results overall so it is the one that is chosen to be used in the next calculations. The fact that LSDLOP24, LSATDOP24 and DLOP are lasts reflects that 50 sensors are not enough for this case study network.

VI-2 Source Identification Scenario and Contamination Extend Estimation

A simulation of a contamination, on the node S of case study network at noon for 2 hours, is used to create the sensor responses. Those are proceeded by the source identification algorithm to get the potential sources of contamination and their ranking at their evaluated times of contamination. The source identification results are then used to estimate the potential propagation of the contamination.

First, the simulation of the contamination is done with a boundary condition of 1 mg/L at node S between 36h and 38h.

Then, the source identification results are given in Table 19. Depending on the observation time, different results are found. The higher the observation time, the smaller the rank but the bigger is the potential source number. As time passes, there is more positive sensor responses and therefore the number of potential source of contamination increases. However, as we get more information, the real source of contamination is better estimated, and therefore the rank decreases.

Table 19: Source identification results for case network and different OT.

OT	Rank	Potential souce node number
2h	125	131
4h	29	1204
8h	29	2906

Finally, there is the estimation of the propagation. The source identification algorithm not only gives the list of potential sources of contamination but also the times of potential contamination. The computation of the propagation is done with as much injection node as potential sources of contamination. Depending on the potential node being source at each time, the injection used is either 0 or a fixed value. That value is taken as the Contribution of that potential node. The more a node contributes to the contamination, the highest is its value for the estimation of the propagation. Therefore the result is not a concentration but some transported value: the dimensionless rank. The highest the value (closer to 1) is for a time and location, the highest the chance is that the network is contaminated at this location. Also, different results are given by the source identification algorithm for different observation times, therefore different results are found when simulating the propagation. The Figure 67 and Figure 68 compare the different propagation estimations with the real propagation. The scales are not the same, the real propagation is given in mg/L (1 mg/L is injected at the inlet) and the estimations with a value between 0 and 1 as given in Table 20.

The results have been filtered to not show the network because of confidential information constrain. That explains the no continuity of the contamination on the figures, which in fact are not actually the case.

Table 20: Scales of figure 69 and 70.

=0 mg/L
>0 mg/L
>0.001 mg/L
>0.01 mg/L
>0.1 mg/L

=0
>0
>0.001
>0.01
>0.1
>0.4
>0.8

The Figure 67 gives the real propagation of the contamination as well as the estimation of the contamination when the source identification has been done with an observation time of 2h. The Figure 68 gives the estimations for both OT = 4h and OT = 8h. As seen in Table 19, the more is the observation time, the higher is the number of potential sources and therefore that explains why the coverage of the network by the estimated contamination gets bigger with the observation time.

The Figure 67 compares the propagation with its estimation with an observation time of two hours. The results have been taken at the times 36h, 40h and 44h, which correspond to 0h, 4h and 8h after the beginning of the contamination injection.

It can be seen that the estimation gives good results with the propagation being part of its coverage. The Figure 68 gives the estimation for OT = 4h and OT = 8h. Those two estimations give a bigger coverage of the network than the estimation with OT = 2h. However, the higher is the observation time, the better is the contamination area drawn by the probabilities intensity. The result of OT = 8h after eight hours of contamination gives really good estimation of the contamination (in dark red).

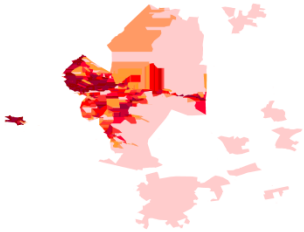
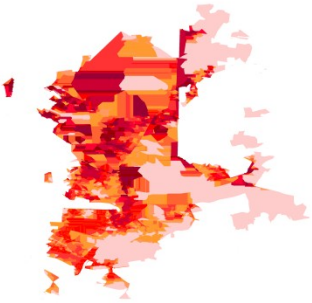
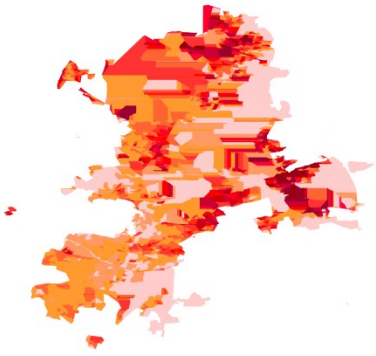
	Contamination	Estimation OT = 2h
0h		
4h		
8h		

Figure 67: Comparison Contamination and evaluation OT = 2h for time 36h, 40h and 44h.

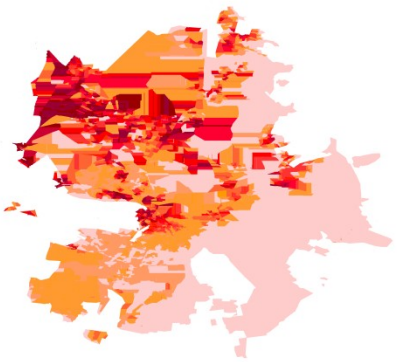
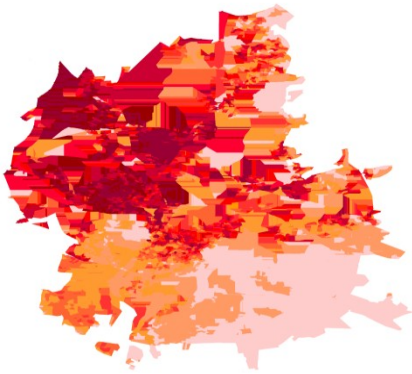
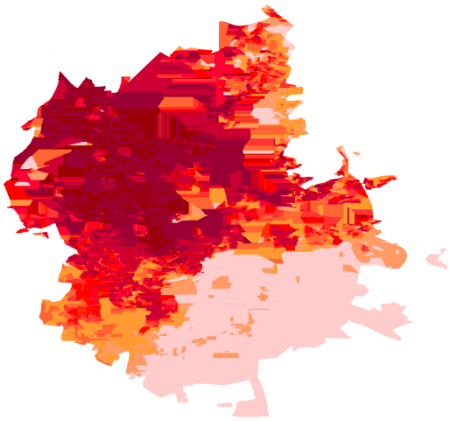
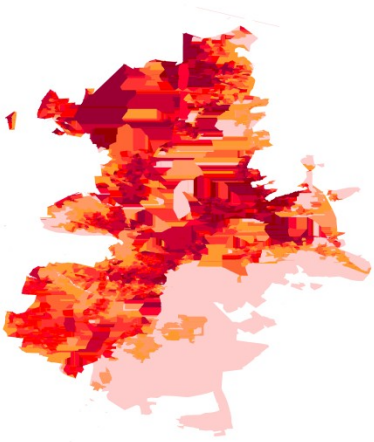
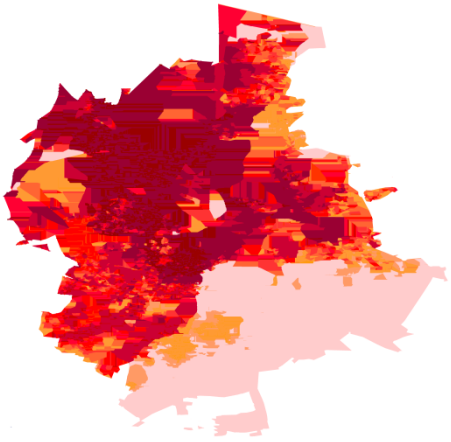
	Estimation OT = 4h	Estimation OT = 8h
0h		
4h		
8h		

Figure 68: Comparison evaluation OT = 4h and OT = 8h for time 36h, 40h and 44h.

Afterwards, we want to test if the imperfect mixing and dispersion enhancements of the transport model influences the estimation of the propagation.

VI-3 Imperfect Mixing

In the case study network, there are more than 100 double T-junctions. However, only one double T-junction has been found with equal diameters, and 15 when diameters are not equals, which are in the right hydraulic conditions (i.e., 2 inflows and 2 outflows) as defined in chapter II. The law found in chapter II is generalized to all diameters (Reynolds number analogy) to see if those can affect the transport of chemical agent. The extrapolation of the function needs to be justified in further researches.

Imperfect mixing results show that the differences from usual transport simulation are limited to two small regions of tens of node. Still, in those areas the error can be superior to 20%. Imperfect mixing is mostly negligible in the two networks that have been studied. However, it has been shown from the experiments that the imperfect mixing at double T-junction have an impact on the transport of contaminants. Any network should be tested to see if it has an actual impact.

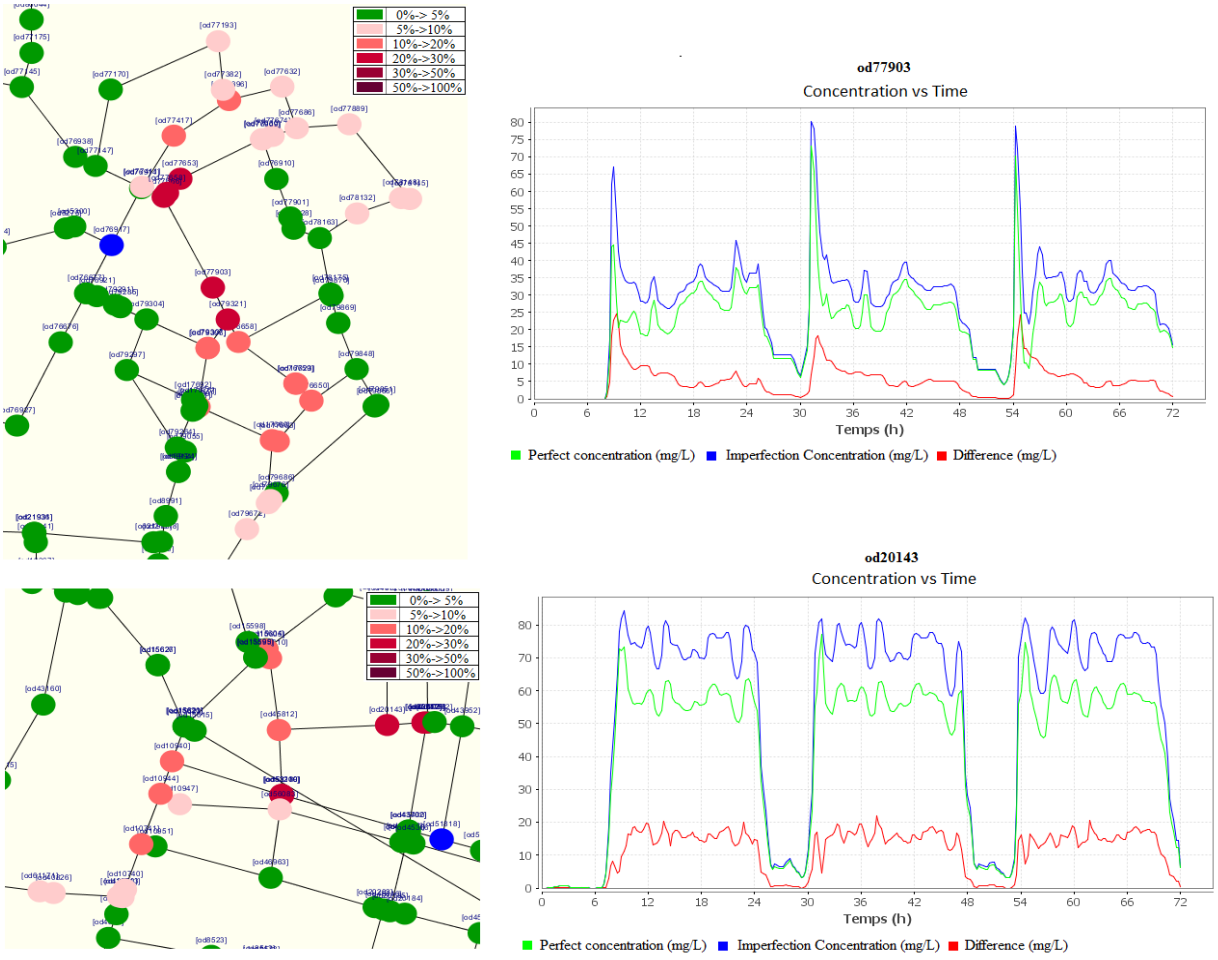


Figure 69: Perfect and imperfect mixing comparison on test case network.

VI-4 Dispersion

In this part the results given by the model with dispersion are compared with the model without it for the contamination at node S. The results are given in Figure 70 with the no dispersion propagation on left at 4h after the beginning of the contamination and the absolute difference to it with the dispersion effect included propagation. The scales are given in Table 21.

Table 21: Scales for figure 72.

=0 mg/L
>0 mg/L
>0.001 mg/L
>0.01 mg/L
>0.1 mg/L

=0 mg/L
>0 mg/L
>0.01 mg/L
>0.05 mg/L
>0.1 mg/L

Figure 70: Dispersion comparison at 4h. real propagation on left, absolute difference with dispersion on right.

The interpretation of the results is that the dispersion does not change much the coverage of the contamination as can be compared the Figure 67 and the Figure 70 (left) at 4h. However the difference value from the models with and without dispersion can be superior to 10% of the initial contamination concentration. When analyzing the whole simulation (Figure 71), most part of the network present significant differences from the no dispersion transport model.

Figure 71: Dispersion difference to real propagation for the whole simulation.

Two characterizations of the dispersion can be seen that explain that results. The first is due to a difference in time of propagation and in this case is mostly present near the contamination source. The second type is when a peak is dispersed, the peak value is lower and is diffused in time, it is shown in Figure 72. The further in the network the node is from the contamination source, the more chance that effect is to appear.

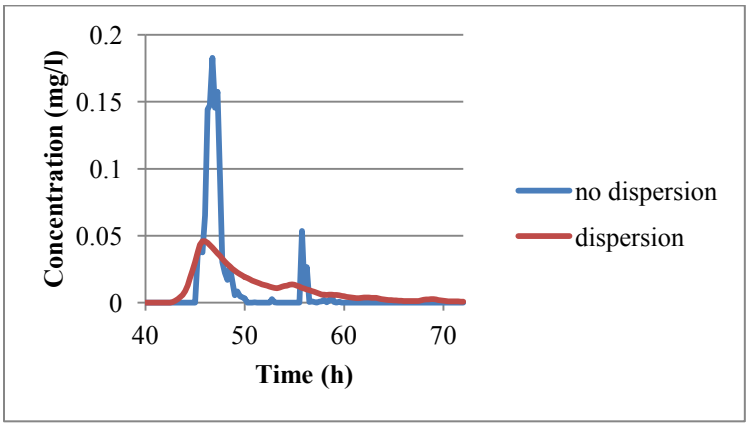


Figure 72: Dispersion effect example.

The dispersion does not need to be used to estimate the propagation of a contamination as the front is the same in this case. However it should be taken into account when transporting chemical agent and that we want to know its concentration.

GENERAL CONCLUSION

The operators are currently installing real time sensors in order to better manage and secure the water distribution networks. Those facilities are vulnerable and critical infrastructures that require protection against accidental and intentional contamination in the network.

The aim of this thesis is to develop different methods that allow the accurate identification of the source of contamination as well as the estimation of its propagation. Different methods have been proposed to complete current methods in three aspects: transport model, source identification evaluation and optimal sensor placement.

The transport model is enhanced in order to have a good model to perform the estimation of the propagation. Indeed, current models do not take into account the two phenomena that are the imperfect mixing and the dispersion. The first one has recently been extended to cross junctions. The prevalence of double T-junctions with respect to simple cross junctions in the French water distribution networks was one of the motivation of studying the imperfect mixing. The dispersion effect concerns the modeling of the three dimensional dynamics which are caused by the velocity profile and the radial diffusion. That effect may appear both on laminar and turbulent regimes. Current transport models are one dimensional advection reaction models that only use the average velocity inside the pipe. In this thesis, we have focused on the effect on the laminar regime, postponing the study of the turbulent regime to future works. Those enhancements to the transport model are then tested on a large network test case to evaluate their benefits.

The transport model considered here is composed of two parts: transport of chemical agent inside pipes on the one hand and mixing at junctions on the other hand. Since the most dangerous scenario is the case of the conservation transport, only this case has been considered. The study of the possible reaction with the biofilm is thus postponed to future developments. Concerns exist on imperfect mixing at structures such as crosses and double T-junctions, and could affect transport simulation results. Cross studies have been done and showed that it can have consequences on the estimation of the propagation of contaminants. Depending on the Reynolds number at each inlets and outlets, the mixing may not be perfect. In some cases, two separate flows appear at the intersection hindering the mixing. The classic model is in-between the perfect mixing and the non mixing state. For double T-junctions, depending on the distance between the two T-junctions, the mixing may also present imperfect mixing like in the cross case. In this thesis, the study has been done as follow. Large numbers of CFD simulations have been computed, direct numerical simulations for laminar cases and large eddy simulations for turbulent cases. A four parameter table is filled with the results of those simulations and the Kriging interpolation is used to complete the entries that are not in the table. A Delaunay triangulation

method is used in the design of experiment phase to select the best design points to simulate in order to have as much information as possible. As the simulation cost is high, this avoids too many simulations that would be needed in a structured design. A 1-D-law is finally created to take into account every configuration of flow rates and concentrations possible. The results are that when the velocity of the straight inlet is smaller than the velocity at the second inlet, the mixing is almost perfect. Otherwise, the mixing is imperfect and depends on the outlet flows as well as the distance between the T-junctions. The higher the distance is, the more the mixing is close to perfect. The same applies for the Reynolds number, the higher it is, the less there is imperfect mixing. However, contrary to the distance inter-T-junctions, it does not necessarily converge to the perfect mixing but to a constant state of imperfect mixing. It has been compared to experiments: the laminar case was hindered by experimental measurement issues such as difficulties in measuring low velocity. However, turbulent CFD simulations have the same behavior as experiments, thus validating the model. In the turbulent case, the CFD simulations may induce a little more imperfect mixing, this may be due to some turbulence being neglected. Simulations and experiments have been carried out using the real geometry, that takes into account physical matters such as joints, but it did not change the results. Finally, an imperfect mixing 1D-law for double T-junction has been created that uses a CFD based data table and a high order interpolation Kriging. It completes the precedent studies of experiments on cross and double T-junctions.

Aside from mixing simplification, the current transport model may also neglect the dispersion effect. Current models are for now one dimensional and only use the average velocity inside the pipe. The dispersion model proposes an enhanced version of that model that takes into account velocity profile and radial diffusion. It has been studied here under the laminar regime with a Poisson profile. A backtracking scheme has been developed on water distribution networks in order to model the transport. That permits to concentrate the information at the nodes instead of the pipes. The velocity profile can then be taken into account, and the velocity value is different on each radius layer of the pipe. An integral is done to give the average results of concentration at the nodes. Additionally, a random walk algorithm is coupled to it in order to simulate the radial diffusion. A Gaussian distribution is both used to get spatial steps on x-axis and y-axis with symmetric boundary. It is then compared to CFD simulations and gives similar results. A small application of the algorithm on a two pipe networks shows that depending on the diffusion coefficient, the results change. If the coefficient is small, the result is close to the velocity profile backtracking results, if it is big, it is close to the one dimensional no dispersive model results. Finally, experiments have been made and measurements have shown a good matching with the model proposed of backtracking and random walk. The trends are the same, only times differ. The experiments give results that are in advance, it may be due to sensors being at the center of the pipes. It also shows that the dispersion is also present in turbulent case, and more research in that subject is necessary. Finally, a model of backtracking, for axial velocity, is coupled to a random walk, for radial diffusion, to model the phenomena of dispersion. It gives

similar concentration results as previous studies, using a virtual diffusion term, and extends them for low diffusion coefficients.

Additionally, to the establishment of an accurate transport model, a source identification method and a sensor placement algorithm are needed in case of accidental or intentional contaminant injection inside the network. Taking into account contamination responses at sensors, the contamination source needs to be found in order to turn off its input. It is also essential to simulate the propagation of the past contamination as well as its current situation. The source identification methods are well studied in the literature. The method proposed here is coupled to two new sensor placement methods to get a solution dedicated to source identification. A probabilistic propagation model is then described; it uses a source identification potential nodes list as well as their rankings.

The aim of the source identification algorithm is to get the list of potential sources of contamination and to rank them according to the probability of being the true source of contamination. The method used is a backtracking algorithm on both positive and negative responses at sensors. It traces back the possible sources of each response. Then it ranks them on the one that explains most of the positive responses. The rank given by the algorithm to the true source of contamination node is then used to create a criterion named Contribution whose value is between 0 and 1. A Monte Carlo method is used to evaluate a sensor placement on its capacity in performing a good source identification. Four criteria are computed: the average time to detection, the detection likelihood, the accuracy and the specificity. Finally, a backtracking is applied on positive and negative answer to create the binary input/output matrix of contamination. It extends previous studies to large network and defines new criteria to evaluate the source identification performance of sensor placement.

The results of the source identification depend on the sensor placement it is applied on. Eight method/parameter configurations have been tested in this thesis. The first twos have been given by a greedy algorithm developed by Piller et al. (2015) : one applied on average time to detection criterion and the other one based on detection likelihood. Another twos are also coming from a greedy algorithm but with the Contribution criteria and a backtracking time of 4h and 24h respectively. Last four ones are given by the local search on graph algorithm applied to the first two sensor placements (ATDOP and DLOP) as initialization and backtracking time of 4h and 24h respectively. A result is that the Contribution factor may be too dependent on the detection likelihood criterion. Moreover, the backtracking time influences directly the average time to detection: the smaller it is the smaller the average time to detection is. Finally, the greedy algorithms adapted to source identification, with Contribution criteria, give better results overall, considering average time to detection, detection likelihood, non accuracy and specificity. However, it is too computationally demanding to be used on large networks. The local search algorithm is giving good results and is much quicker. It can even be used on large networks. Finally, two new methods of sensor placement have been developed. One is the greedy algorithm and the other is the local search on graph. Both are applied to the source identification

Contribution criterion. They give better result overall for source identification but take much longer time.

The methods have been applied on a large test case network. The sensor placement giving better results overall is the local search algorithm with average time to detection sensor placement initialization and backtracking time of 4h. The source identification algorithm is then used to get the list of potential sources of contamination and their rankings. The dimensionless value of that ranking, Contribution between 0 and 1, is then used to estimate the propagation inside the network. For every time step, every potential sources of contamination concentration is either given the value equals to its Contribution, when it is potentially contaminated at that time, or zero. The estimate propagation matches the real propagation, and the higher the time of observation, the more accurate the model is. Finally, the two enhancements of the transport model have been tested. Fifteen double T-junctions, not constrained to equal diameters, are in hydraulic conditions to be affected by imperfect mixing. However, overall, only very small regions present differences and it may be neglected for that network. Concerning the dispersion effect, small differences have been observed when comparing the coverage of the propagation. However, the value of the concentration of the contamination may be greatly affected and can be easily found in the network with deviation more than ten percent of the injected concentration.

As a conclusion, local search on graph is enhancing the capability of a sensor placement to identify the source of contamination. A small backtracking time may be used to induce small average time to detection, which is important in case of contamination breakthrough. This capability can then be evaluated with a Monte Carlo method on the four parameters: average time to detection, detection likelihood, probability of missing the true source of contamination and ranking on detected scenario. In case of contamination, the source identification algorithm is then used to give the ranked potential list of contaminations. Afterwards, that list is put as injection point inside the network to evaluate the propagation. The higher the observation time, the best the model is estimating the contamination location. Finally, the imperfect mixing at double T-junctions may be tested beforehand on the network to know if it can affect the transport. And the dispersion effect has mostly an influence on the value of the concentration. The coverage of the contamination stays more or less the same. That may depend on the hydraulic regime of the network, as turbulent regime dispersion effect is underestimated by the model proposed.

Next researches can focus on imperfect mixing for double T-junctions with unequal diameters. The dispersion model differences on propagation results can be studied to see its effect on sensor responses calculation, and therefore sensor placement. An extended statistical study on the evaluation of source identification performance can be performed. And multi-objective frameworks can be used for sensor placement to draw Pareto front on the source contamination criteria defined. Finally, imperfect sensor study needs to be done on the source identification algorithm and negative responses use need careful attention as it is directly related to contamination detection threshold and false negative rate.

Bibliography

Abraham, W. 1943. *On the efficient Design of Statistical Investigations.* s.l. : The Annals of Mathematical Statistics, 1943. Vol. 14, No. 2 (Jun., 1943), pp. 134-140.

Archambeau, F., Méchitoua, N. and Sazik, M. 2015. *Code Saturne: A Finite Volume code for the computation of turbulent incompressible flows – Industrial Applications.* s.l. : International Journal on Finite Volumes, Episciences.org, 2015.

Bergmann, M., Iollo, A. and Telib, H. 2012. *FFAST Future Fast Aeroelastic Simulation Technologies D2.17: Implementation and validation of hybrid basis fusion ROM method.* 2012.

Berry, J., Hart, W. E., Phillips, C. A., Uber, J. G., Watson, J.-P. 2006. *Sensor Placement in Municipal Water Networks with Temporal Integer Programming Models.* s.l. : Journal of Water Resources Planning and Management, 2006. Volume 132, SPECIAL ISSUE:Drinking Water Distribution Systems Security.

Boulos, P., Altman, T., Jarrige, P.-A., Collevati, F. 1994. *An event-driven method for modelling contaminant propagation in water networks.* s.l. : Appl. Math. Modelling,, 1994. Vol. 18, February.

Braun, M., Bernard, T., Ung, H., Piller, O., Gilbert, D. 2014. *Model based investigation of transport phenomena in water distribution networks for contamination scenarios.* s.l. : Procedia Engineering - 12th International Conference on Computing and Control for the Water Industry, CCWI2013, 2014. Volume 70, 2014, Pages 191–200.

Carlier, M. 1980. *Hydraulique générale et appliquée.* 1980.

Chauvet, P., Galli, A. 1982. *Universal kriging.* s.l. : Centre de géostatistique - Ecole des Mines de Paris, 1982.

Choi, C. Y., Shen, J. Y. and Austin, R. G. 2008. *Development of a Comprehensive Solute Mixing Model (AZRED) for Double-Tee, Cross, and Wye Junctions.* s.l. : Water Distribution Systems Analysis , 2008. 1-10.

Colebrook, C. F. 1939. *Turbulent flow in pipes, with particular reference to the transition region between smooth and rough pipe laws.* s.l. : Journal of the ICE - Proceedings of the Institution of Civil Engineers, 1939. Volume 11 Issue 4, FEBRUARY 1939, pp. 133-156.

Conn, A. R., Gould, N. I. M. and L., Toint Ph. 1988. *Global Convergence of a Class of Trust Region Algorithms for Optimization with Simple Bounds.* s.l. : SIAM J. Numer. Anal, 1988. 25(2), 433–460.

Constans, S., Bremond, B. and Morel, P. 2003. *Simulation and control of chlorine levels in water distribution networks.* s.l. : Procedia Engineering - 16th Water Distribution System Analysis Conference, WDSA2014 — Urban Water Hydroinformatics and Strategic Planning, 2003. Volume 89, 2014, Pages 239–246.

Costa, J.-P., Pronzato, L. and Thierry, E. 1999. *a comparison between kriging and radial basis function networks for nonlinear prediction.* s.l. : IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing (NSIP'99), Antalya, Turkey, June 20-23, 1999.

Cressie. 1993. *Statistics for spatial data.* 1993. ISBN: 978-0-471-00255-0.

Dawsey, W., Minsker, B. and VanBlaricum, V. 2006. *Bayesian Belief Networks to Integrate Monitoring Evidence of Water Distribution System Contamination.* s.l. : Journal of Water Resources Planning and Management, 2006. Volume 132, SPECIAL ISSUE:Drinking Water Distribution Systems Security.

De Sanctis, A., Shang, F. and Uber, J. 2010. *Real-Time Identification of Possible Contamination Sources Using Network Backtracking Methods.* s.l. : Water Resour. Plann. Manag., 2010. ASCE 136 (4). 444-453..

Deuerlein, J., Piller, O. and Montalvo, I. A. 2014. *Improved Real-time Monitoring and Control of Water Supply Networks by Use of Graph Decomposition.* s.l. : Procedia Engineering, 2014. 2014, 89, pp.1276-1281..

Epanet. 2000. *EPANET 2 USERS MANUAL.* 2000. EPA/600/R-00/057.

Fabrie, P., Gancel, G., Mortazavi, I., Piller, O. 2010. *Quality modeling of water distribution systels using sensitivity equations.* s.l. : journal of hydraulic engineering-ASCE, 2010. 2010, 136 (1), p. 34 - p. 44.

Gill, W. N. and Sankarasubramanian, R. 1970. *Exact analysis of unsteady.* s.l. : Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, 1970. Vol. 316, No. 1526 (May 5, 1970), pp. 341-350.

Godsil, Chris and Royle, Gordon. 2004. *Algebraic Graph Theory.* s.l. : Springer. Sect. 6.8., 2004.

Guan, J., et al. 2006. *Identification of Contaminant Sources in Water Distribution Systems Using Simulation Optimization Method: Case Study.* s.l. : Journal of Water Resources Planning and Management, 2006. 07/2006; 132:252-262. DOI: 10.1061/(ASCE)0733-9496(2006)132:4(252).

Ho, C.K. and Khalsa, S.S. 2008. *EPANET-BAM: Water Quality Modeling with Incomplete Mixing in Pipe Junctions.* s.l. : Journal Hydraulic Engineering, 2008. 134(9), 1236-1244. (SAND2008-0166J).

Jun, Yin, Szu, Hui Ng and Kien, Ming Ng. 2009. *a study on the effects of parameter estimation on kriging model's prediction error in stochastic simulations.* s.l. : Simulation Conference (WSC), Proceedings of the 2009 Winter, 2009. 978-1-4244-5770-0.

Kay, M. D., Beckman, R. J. and Conover, W. J. 1979. *A Comparaison of three Methods for Selecting Values of Input Variables in the Analysis of Output from a computer Code.* s.l. : Technometrics, 1979. Vol. 42, No. 1, Special 40th Anniversary Issue (Feb., 2000), pp. 55-61.

Krause, Andreas, Leskovec, Jure, Isovitsch, Shannon, Xu, Jianhua, Guestrin, Carlos, VanBriesen, Jeanne, Small, Mitchell, Fischbeck, Paul. 2006. *optimizing sensor placements in water distribution systems using submodular function maximization.* s.l. : PROCEEDINGS - Water Distribution Systems Analysis Symposium 2006, 2006. 10.1061/40941(247)109.

Laird, C. D., Biegler, L. T. and Van Bloemen Waanders, B. G. 2006. *Mixed-integer approach for obtaining unique solutions in source inversion of water networks.* s.l. : Journal of Water Resources Planning and Management, 2006. Special Issue on Drinking Water Distribution Systems Security 132, 242 (2006).

Lee, B. H. and Deininger, R. A. 1992. *Optimal Locations of Monitoring Stations in Water Distribution System.* s.l. : Procedia Engineering - Computing and Control for the Water Industry (CCWI2015) Sharing the best practice in water management, 1992. Volume 119, 2015, Pages 1308–1317.

Lee, Y. 2004. *Mass dispersion in intermittent laminar flow.* s.l. : PhD, University of Cincinnati, Engineering : Environmental Engineering., 2004.

Li, Z. 2006. *Network Water Quality Modeling with Stochastic Water Demands and Mass Dispersion.* s.l. : PhD, University of Cincinnati, Engineering : Environmental Engineering, 2006.

Liu, L., Sankarasubramanian, A. and Ranjithan, S. R. 2011. *Logistic regression analysis to estimate contaminant sources in water distribution systems.* s.l. : Journal of Hydroinformatics - & IWA Publishing 2011, 2011. 13.39 545.

Liu, Shuming and Auckenthaler, Pierre. 2014. *Optimal sensor placement for event detection and source identification in water distribution networks.* s.l. : Aqua, 2014. 02/2014; 63(1):51.

Neupauer, R. 2011. *Adjoint Sensitivity Analysis of Contaminant Concentrations in Water Distribution Systems.* s.l. : Journal of Engineering Mechanics, 137(1), 31-39., 2011.

Ostfeld, Avi, Uber, James G., Salomons, Elad, Berry, Jonathan W., Hart, William E., Phillips, Cindy A., Watson, Jean-Paul, Dorini, Gianluca, Jonkergouw, Philip, Kapelan, Zoran, Pierro, Francesco di, Khu, Soon-Thiam, Savic, Dragan, Eliades, Demetrios, et al. 2008. *The Battle of the Water Sensor Networks (BWSN): A Design Challenge for Engineers and Algorithms.* s.l. : Journal of Water Resources Planning and Management-asce, 2008. 10.1061/(ASCE)0733-9496(2008)134:6(556).

Perelman, L. and Ostfeld, A. 2010. *Bayesian Networks for Estimating Contaminant Source and Propagation in a Water Distribution System Using Cluster Structure.* s.l. : Water Distribution Systems Analysis 2010 - PROCEEDINGS, 2010. 10.1061/41203(425)40.

Piller, O. and Tavard, L. 2014. *Modeling the transport of physicochemical parameters for water network security.* s.l. : Procedia Engineering, 70, 1344-1352., 2014.

Piller, O. 1995. Modeling the behavior of a network - Hydraulic analysis and sampling procedures for parameter estimation. *PhD thesis.* s.l. : Applied Mathematics thesis from the University of Bordeaux (PRES), defended on, 288 pages, Talence, France., 1995.

Piller, O., Deuerlein, J., Gilbert, D., Weber, J-M. 2015. *Installing fixed sensors for double calibration and early-warning detection purposes.* s.l. : Computing and Control for the Water Industry (CCWI2015) Sharing the best practice in water management, 2015. Volume 119, 2015, Pages 564–572.

Porteau. [Online] [Cited: November 24, 2015.] http://porteau.irstea.fr/.

Preis, A. and Ostfeld, A. 2006. *multiobjective sensor design for water distribution systems security.* s.l. : Water Distribution Systems Analysis Symposium 2006 - PROCEEDINGS, 2006. 10.1061/40941(247)107.

Preis, A. and Ostfeld, A. 2006. *Optimal Sensors Layout for Contamination Source Identification in Water Distribution Systems.* s.l. : Water Distribution Systems Analysis Symposium 2006 - PROCEEDINGS, 2006. 10.1061/40941(247)127.

Propato, M. and Piller, O. 2006. *Battle of the Water Sensor Networks.* s.l. : 8th annual Water Distribution System Analysis Symposium, 2006. August 27-30 2006. 8 p..

Propato, M., Sarrazy, F. and Tryby, M. 2010. *Linear Algebra and Minimum Relative Entropy to Investigate Contamination Events in Drinking Water Systems.* s.l. : Journal of Water Resources Planning and Management, 2010. Volume 136, Issue 4 (July 2010) 10.1061/(ASCE)WR.1943-5452.0000059.

Propato, M., Tryby, M. E. and Piller, O. 2007. *Linear Algebra Analysis For Contaminant Source Identification in Water Distribution Systems.* s.l. : World Environmental and Water Resources Congress 2007, 2007. Tampa, Florida, USA, 15-19 May 2007.

Rathi, S. and Gupta, R. 2014. *Sensor placement methods for contamination detection in water distribution networks: a review.* s.l. : 16th Water Distribution System Analysis Conference, WDSA2014 — Urban Water Hydroinformatics and Strategic Planning, 2014. Volume 89, 2014, Pages 181–188.

Romero-Gomez, P., Li, Z., Choi, C. Y., and Buchberger, S. G. 2009. *Axial dispersion coefficients for laminar flows in water distribution systems.* s.l. : World Environmental and Water Resources Congress 2009 - PROCEEDINGS, 2009. 10.1061/41036(342)7.

Seth, A., et al. 2016. *Testing Contamination Source Identification Methods for Water Distribution Networks.* s.l. : Journal of Water Resources Planning and Management, 2016. 04016001.

Shang, F., Uber, J. G. and Polycarpou, M. M. 2002. *Practical back-tracking algorithm for water distribution system analysis.* s.l. : Journal of Environmental Engineering, 2002. 128(5), pp. 441 -.

Sir 3S. [Online] [Cited: November 24, 2015.] http://www.3sconsult.de/sir3s.html.

SMaRT-Online^{WDN}. [Online] [Cited: November 24, 2015.] http://www.smart-onlinewdn.eu/.

Taylor, G. 1953. *Dispersion of soluble matter in solvent flowing slowly.* s.l. : Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, 1953. Vol. 219, No. 1137. (Aug. 25, 1953), pp. 186-203.

Todiny, E. and Pilati, S. 1987. *A gradient algorithm for the analysis of pipe networks.* s.l. : Conference: Proc. Computer Applications for Water Supply and Distribution, 1987.

Tryby, M. E., Propato, M. and Ranjithan, S. R. 2010. *Monitoring Design for Source Identification in Water Distribution Systems.* s.l. : Journal of Water Resources Planning and Management, 2010. Volume 136, Issue 6 (November 2010) 10.1061/(ASCE)WR.1943-5452.0000080.

Ung, H., Piller, O. and Gilbert, D. 2014. *Quasi-real time modeling for security of a drinking water distribution network.* s.l. : Procedia Engineering, 2014. vol. 70, p. 800 - 809.

Waeytens, J., Chatellier, P. and Bourquin, F. 2015. *Inverse computational fluid dynamics: influence of discretisation and model errors on flows in water network including junctions.* s.l. : ASME Journal of Fluids Engineering, 2015. 137,9..

Wagner, D., Neupauer, R. M. and Cichowitz, C. 2015. *Adjoint-Based Probabilistic Source Characterization in Water-Distribution Systems with Transient Flows and Imperfect Sensors.* s.l. : Journal of Water Resources Planning and Management, 2015. Volume 141, Issue 9 (September 2015) 10.1061/(ASCE)WR.1943-5452.0000508.

Wim, C.M., van Beers, J. and Kleijnen, P.C. 2004. *kriiging interpolation in simulation: a survey.* s.l. : Proceedings of the 2004 Winter Simulation Conference, 2004. pp 115-121.

Yang, C-S., Kao, S-P., Lee, F-B., Hung, P-S. 2004. *twelve different interpolation methods: a case study of SURFER 8.0.* s.l. : XXth ISPRS Congress, 2004. Vol. Technical Commision II.

Acknowledgments

We wish to acknowledge the three following institutes for letting us use their computation cluster to carry out the simulations performed in this project:

- Mésocentre de Calcul Intensif Aquitain (MCIA), Avakas supercomputer
- Centre Informatique National de l'Enseignement Supérieur (CINES), Jade supercomputer
- Irstea Clermont computation grid

The research part of the *SMaRT-Online*^{WDN} project is supported by the German Federal Ministry of Education and Research (BMBF; project: 13N12180) and by the French Agence Nationale de la Recherche (ANR; project: ANR-11-SECU-006).

Appendices

Appendix A: Transport model

The transport model aim is to model the propagation of a chemical agent inside the network. One use is to model the transport of chlorine inside water distribution network. One other is to compute the residence time of water. Still one other application can be to model the thermal state inside the distribution network.

A-I Model

A-I-1 Transport Equation

The usual equation used for the transport of chemical product in a water distribution network is the advection reaction 1-D equation:

$$\frac{\partial C}{\partial t} + u\frac{\partial C}{\partial x} + KC^{\alpha} = 0,$$

with u the velocity, C the concentration, K and α the kinetic constant and order of the reaction.

Additionally, mixing condition at nodes is generally considered as perfect:

$$\forall \text{ noeud } i, C_{out(i)} = \frac{\sum q_{in(i)} C_{in(i)}}{\sum q_{in(i)}},$$

with in(i) the pipes whose flow going into node i.

The thermal equation is a little different (except for $T_{ext} = 0$):

$$\frac{\partial T}{\partial t} + u\frac{\partial T}{\partial x} + K_T(T - T_{ext}) = 0,$$

with T the temperature, K_T the thermal constant and T_{ext} the external temperature.

A-I-2 Transport Graph

The transport model is applied on a non-oriented network like the example given Figure 73. The chemical agent is injected at boundary condition nodes such as resources or injection points. It's transport depends on the water velocity inside pipes.

The pipe model is composed of:

- Length (m)
- Diameter (mm)
- Velocity(m/s) (that depends on time)

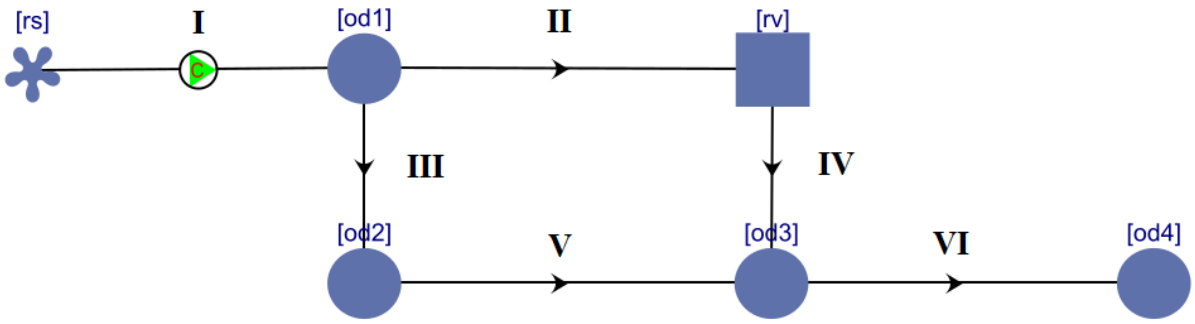


Figure 73: Water distribution network example.

Also three hydraulic types of node exist depending on their role in the water distribution system:

- ordinary: either demand node or pipe junction
- tank: model a water tank
- resource: source of water with constant head, concentration, age…

Finally, water quality types can be defined to represent different boundary conditions. Each condition is classed in one of the four quality types:

- ordinary: mixing node (ordinary node)
- fixed limit condition: fixed concentration function on time (ordinary, tank and resource nodes)
- injection node: mixing with additional amount of water at a fixed concentration (ordinary node)
- initial condition: tank with initial amount of water and concentration (tank node)

A-II Model Types

Different models exist, some are time-driven, and others are event-driven. They use the same equation presented previously but used different algorithms and schemes to solve them.

A-II-1 Hybrid Porteau Model

It is a time-driven model, it is based on a discretization of the calculation time on quality time step. That time step is generally a divisor of the hydraulic time step to easier the calculation.

The model uses a hybrid Eulerian scheme that is composed of two parts: one method of characteristics followed by a Crank-Nicholson scheme. It is applied on a segment divided into discretization $dx = floor(\frac{Length}{u*dt})$. The chemical agent is first transported on $dt_{moc} = \frac{dx}{u}$ by method of the characteristics. (see Fabrie et al. (2010))

$$\forall(x,t), C(x,t) = C_0(x - u \times t).$$

Then a θ–scheme is applied on the time left dt = dt_{qua} - dt_{moc}.

$$C(x + dx, t + dt_{qua}) = C(x + dx, t + dt_{moc}) - h * \left(\theta\left(C(x + dx, t + dt_{moc}) - C(x, t + dt_{moc})\right) - (1 - \theta)C(x, t + dt_{qua})\right),$$

$$h = \frac{u \times dt}{dx}.$$

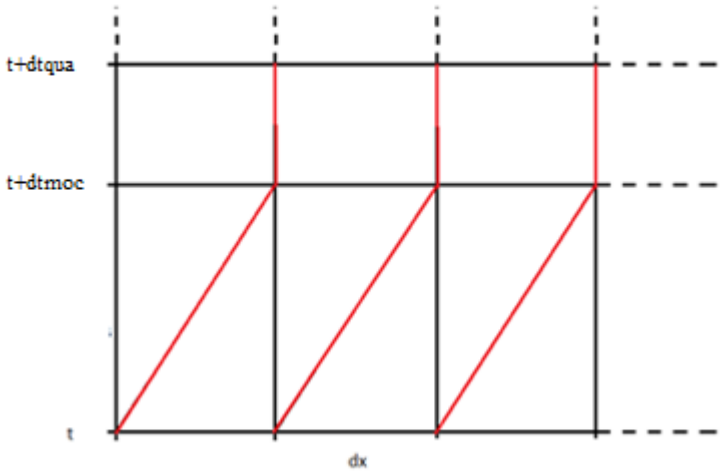


Figure 74: Hybrid Eulerian method.

The Eulerian model asset is that the scheme can easily be changed, increasing the order if needed. Also more phenomena such as diffusion can be easily added to the model. The drawbacks are that it needs some memory capacity as it needs to save data on each pipe discretizations. Also interpolation errors may influence the calculation results. However Fabrie et al. (2010) shows that is still better than any 2nd order scheme.

A-II-2 Lagrangian Epanet Model

It is an event-driven model that happens at each change in the boundary condition and when a marker reach the end of a pipe.

The model uses a forward Lagrangian scheme. The quality is transported inside the pipe between two markers as long as no event happens. When limit conditions change, markers are created for every pipe with water going out of this source. When a marker reaches the end of a pipe, the mixing is done and new markers are set to outflowing pipes.

The assets are that the transport is exactly done and that the accuracy only depend on the hydraulic time step and boundary conditions. The drawback comes from the mixing conditions, a limit on the distance between two markers is needed or else too much markers are created and the calculation time exponentially grows.

A-II-3 Forward Adjoint Model (used for dispersion)

This is a new model proposed to solve the transport. It is based on particle backtracking scheme. It is also a time-driven model. For every time step, quality at every node is computed and added to precedent ones into a quality matrix. This matrix is used to calculate future values through a backtracking scheme. The quality at one end of the pipe can be traced back to its origin. The time it has needed to travel through the pipe from one end of the pipe is calculated. The quality value of the source node at that time is given by the matrix with an interpolation method. Then reaction is done accordingly to time passed inside the pipe. Finally mixing on the end node is done to calculate its current value. If the order of the reaction is 1 and the reaction constant is K:

$$C(i) = \begin{cases} [\theta \times C(itime) + (1-\theta) \times C(itime+1)]e^{-K timeinpipe} & \text{si } i \times dtqua - timeinpipe > 0 \\ 0 & \text{sinon} \end{cases},$$

$$\theta = \begin{cases} (i \times dtqua - timeinpipe) - itime \times dtqua & \text{si } itime \neq 0 \\ 1 & \text{sinon,} \end{cases}$$

$$itime = \ floor\left(\frac{i \times dtqua - timeinpipe}{dtqua}\right),$$

with i the current iteration, dtqua the quality time step and timeinpipe the time spend in the pipe computed by the backtracking algorithm

This method asset is that it does not model what is a happening inside the pipe. Also it can be used for the dispersion model presented in this thesis. Its drawbacks are that it needs to store all velocities and quality values of the whole calculation to do next calculation. Also the interpolation precision depends on the quality time step defining the time difference between two values inside the quality matrix.

A-II-4 Summary

More models exist and a non exhaustive list is given Table 22. They are grouped in Eulerian and Lagrangian model as well as direct and adjoint models. The reaction term is not included; usual modeling includes a splitting algorithm to take it into account.

Different Lagrangian models can be used depending on the variables calculated. The forward adjoint model is also a Lagrangian method. It uses the direct formulation of the adjoint scheme applied to do the source identification. Therefore, the same algorithm can be used for both as described on the object part in Appendix D.

Table 22: List of transport models.

	Direct	Adjoint
Equation	$\frac{\partial C}{\partial t} + u\frac{\partial C}{\partial x} = 0$	$\frac{\partial \psi}{\partial \tau} - u\frac{\partial \psi}{\partial \chi} = 0$ with $\boxed{\tau = t_0 - t}$ and $\boxed{\chi = x_0 - x}$ and then $\boxed{d\tau = -dt}$ and $\boxed{d\chi = -dx}$
Eulerian Explicit	$C_{x+dx}^{t+dt} = C_{x+dx}^{t} - u\frac{dt}{dx}\left(C_{x+dx}^{t} - C_{x}^{t}\right)$	$\psi_{\chi-dx}^{\tau-dt} = \psi_{\chi-dx}^{\tau} - u\frac{dt}{dx}\left(\psi_{\chi-dx}^{\tau} - \psi_{\chi}^{\tau}\right)$
Eulerian θ-scheme	$C_{x+dx}^{t+dt} = C_{x+dx}^{t} - u\frac{dt}{dx}[\theta * (C_{x+dx}^{t+1} - C_{x}^{t+1}) + (1-\theta)(C_{x+dx}^{t} - C_{x}^{t})]$ (Porteau)	$\psi_{\chi-dx}^{\tau-dt} = \psi_{\chi-dx}^{\tau} - u\frac{dt}{dx}[\theta(\psi_{\chi-dx}^{\tau-dt} - \psi_{\chi}^{\tau-dt}) + (1-\theta)(\psi_{\chi-dx}^{\tau} - \psi_{\chi}^{\tau})]$
Lagrangian	$C_{x+udt}^{t+dt} = C_{x}^{t}$ (Epanet 1) (Porteau with x=udt)	$\psi_{\chi-udt}^{\tau-dt} = \psi_{\chi}^{\tau}$
Lagrangian	$C_{x+dx}^{t+\frac{dx}{u}} = C_{x}^{t}$ (Epanet 2) $C_{x+dx}^{t} = C_{x}^{t-\frac{dx}{u}}$	$\psi_{\chi-dx}^{\tau-\frac{dx}{u}} = \psi_{\chi}^{\tau}$ (Backtracking with dx=L) $\psi_{\chi-dx}^{\tau} = \psi_{\chi}^{\tau+\frac{dx}{u}}$
Lagrangian	$C_{x+dx}^{t+dt} = C_{x}^{t+dt-\frac{dx}{u}}$ (forward adjoint with dx=L)	$\psi_{\chi-dx}^{\tau-dt} = \psi_{\chi}^{\tau-dt+\frac{dx}{u}}$
Lagrangian	$C_{x}^{t+\frac{dx}{u}} = C_{x-dx}^{t}$	$\psi_{\chi}^{\tau-\frac{dx}{u}} = \psi_{\chi-dx}^{\tau}$

A-III Model Data

In this part are detailed the data needed to carry out the transport computation. Some are fixed such as network data, others are given during the calculation like velocities in pipes.

A-III-1 Initial Data

A-III-1.1 Parameters

Those define the different time step used, the quality time step dividing the hydraulic time step:

- hydraulic time step
- transport time step

A-III-1.2 Network

The graph characteristics are:

- the list of nodes: name, hydraulic and quality types
- the list of pipes: name, length, diameter, begin and end nodes
- the list of double T-junctions: mixing scenario (perfect/imperfect), node 1 and 2, pipe 1 to 5
- the list of crosses: mixing scenario (perfect/imperfect), node, pipe 1 to 4

A-III-1.3 Transport Model

The transport models have been defined previously:

- Hybrid Eulerian
- Forward adjoint
- Mix of Hybrid Eulerian with the Adjoint direct model applied only on laminar pipes

A-III-1.4 Transport Objects

The transport objects define which kind of quality model is used. The concentration model calculates the transport of a reactive chemical agent inside the network. The ageing model gives the residence time. The provenance model gives for every node the provenance of each source defined in the form of a percentage. The thermal model gives the temperature state of the network.

Except for the thermal object, the others use the same equations. Only the parameters may be different, for instance for ageing the reaction constant is fixed at -1 and order at 0.

- Concentration: initial value, mixing law, kinetic constant and order of pipe and tank reaction
- Ageing: initial value, mixing law, ageing boolean of pipes and tanks
- Provenance: index, mixing law
- Thermal: initial value, mixing law, thermal constant of pipes

The initial value defines the value that is fixed on every nodes and pipes at the beginning of the simulation. The mixing law is minimum, average or maximum. The reaction constant and order depends on the pipe or the tank it is applied on. A boolean list is given to the ageing model, the ageing is done inside the pipe only if that boolean is true. It can be used to calculate the time water spends in specific pipes or tanks. The thermal model only needs to be given the thermal constant for each pipes, however extern temperatures are also needed during the calculation phase.

A-III-1.5 Quality Limit Condition Parameters

Those are the list of the different limit condition and their constant parameters.

For fixed limit condition:

- Index

For injection condition:

- Index
- Concentration

For initial condition:

- Index
- Initial concentration
- Initial volume of water

A-III-2 Calculation Data

Those define the data needed at each hydraulic time steps. Quality data are given as a list with size corresponding to the division of the hydraulic time step by the quality time step.

A-III-2.1 Hydraulics

Those are:

- The velocity in pipes
- The new volume of water inside tanks

The flow, needed for the mixing, can be computed from the velocity and the diameter of the pipe. The new volume of water inside tanks is needed to calculate extern inflow or outflow inside the tank.

A-III-2.2 Quality

The limit condition values are given for every time step as to adapt it to real time simulation:

- The fixed limit condition values
- The flows injected

A-IV Main Transport Algorithm

- Here is described the usual transport algorithm for either hybrid Eulerian model or forward adjoint model. The diagram Figure 75 pipes

A-III-1.4 Transport Objects

The transport objects define which kind of quality model is used. The concentration model calculates the transport of a reactive chemical agent inside the network. The ageing model gives the residence time. The provenance model gives for every node the provenance of each source defined in the form of a percentage. The thermal model gives the temperature state of the network.

Except for the thermal object, the others use the same equations. Only the parameters may be different, for instance for ageing the reaction constant is fixed at -1 and order at 0.

- Concentration: initial value, mixing law, kinetic constant and order of pipe and tank reaction
- Ageing: initial value, mixing law, ageing boolean of pipes and tanks
- Provenance: index, mixing law
- Thermal: initial value, mixing law, thermal constant of pipes

The initial value defines the value that is fixed on every nodes and pipes at the beginning of the simulation. The mixing law is minimum, average or maximum. The reaction constant and order depends on the pipe or the tank it is applied on. A boolean list is given to the ageing model, the ageing is done inside the pipe only if that boolean is true. It can be used to calculate the time water spends in specific pipes or tanks. The thermal model only needs to be given the thermal constant for each pipes, however extern temperatures are also needed during the calculation phase.

A-III-1.5 Quality Limit Condition Parameters

Those are the list of the different limit condition and their constant parameters.

For fixed limit condition:

- Index

For injection condition:

- Index
- Concentration

For initial condition:

- Index
- Initial concentration
- Initial volume of water

A-III-2 Calculation Data

Those define the data needed at each hydraulic time steps. Quality data are given as a list with size corresponding to the division of the hydraulic time step by the quality time step.

A-III-2.1 Hydraulics

Those are:

- The velocity in pipes
- The new volume of water inside tanks

The flow, needed for the mixing, can be computed from the velocity and the diameter of the pipe. The new volume of water inside tanks is needed to calculate extern inflow or outflow inside the tank.

A-III-2.2 Quality

The limit condition values are given for every time step as to adapt it to real time simulation:

- The fixed limit condition values
- The flows injected

A-IV Main Transport Algorithm

Here is described the usual transport algorithm for either hybrid Eulerian model or forward adjoint model. The diagram Figure 75 sums up the algorithm.

A-IV-1 Initialization

In the first time all tables are initialized and pipe discretized. First quality values are given with the initial values parameter of quality objects as well as limit conditions.

A-IV-1.1 Nodes

The initial value is given to every node except the initial and fixed limit condition nodes whose have their own value.

A-IV-1.2 Pipes

In case of Eulerian scheme, the pipes are discretized depending on the initial velocity. The discretization used is $dx = floor(\frac{Length}{u \times dt})$. All discretization is given the same initial value.

A-IV-1.3 Non-Oriented Adjacent List

The non-oriented adjacent list is built from the list of pipes. For each node it gives the list of connected pipes. The oriented list that is used when mixing is calculated for every hydraulic time step.

A-IV-2 Calculation

Now for every hydraulic time step, transport is done and results are returned.

A-IV-2.1 Precalculation

First limit condition nodes are updated:

- limit condition nodes are given a new value
- reaction is done on initial condition node

Then, new discretization on pipes is done in function of new velocity as well as interpolation of values.

After that, the reaction is done on pipe with velocity zero for all its discretizations.

Then, balance on flows is done for each node to detect eventual inflows.

Finally, the propagation pipe list is built depending on the velocities direction, going from fixed limit and initial conditions. This is necessary for the hybrid Eulerian model as it is a semi implicit model.

A-IV-2.2 Calculation

Transport is done on pipes in the order of propagation pipe list. For hybrid Eulerian model, first a model of characteristic is done on $dt_{moc} = \frac{dx}{u}$, then a Crank-Nicholson scheme is used on the time left. For forward adjoint model the time and source node is found, value and reaction is given accordingly.

The mixing is done when all incidence pipe transport has been done for that node. Mixing is done given the mixing law: minimum, perfect/average or maximum. Mixing calculation includes inflow pipes, inflows, injection and initial condition.

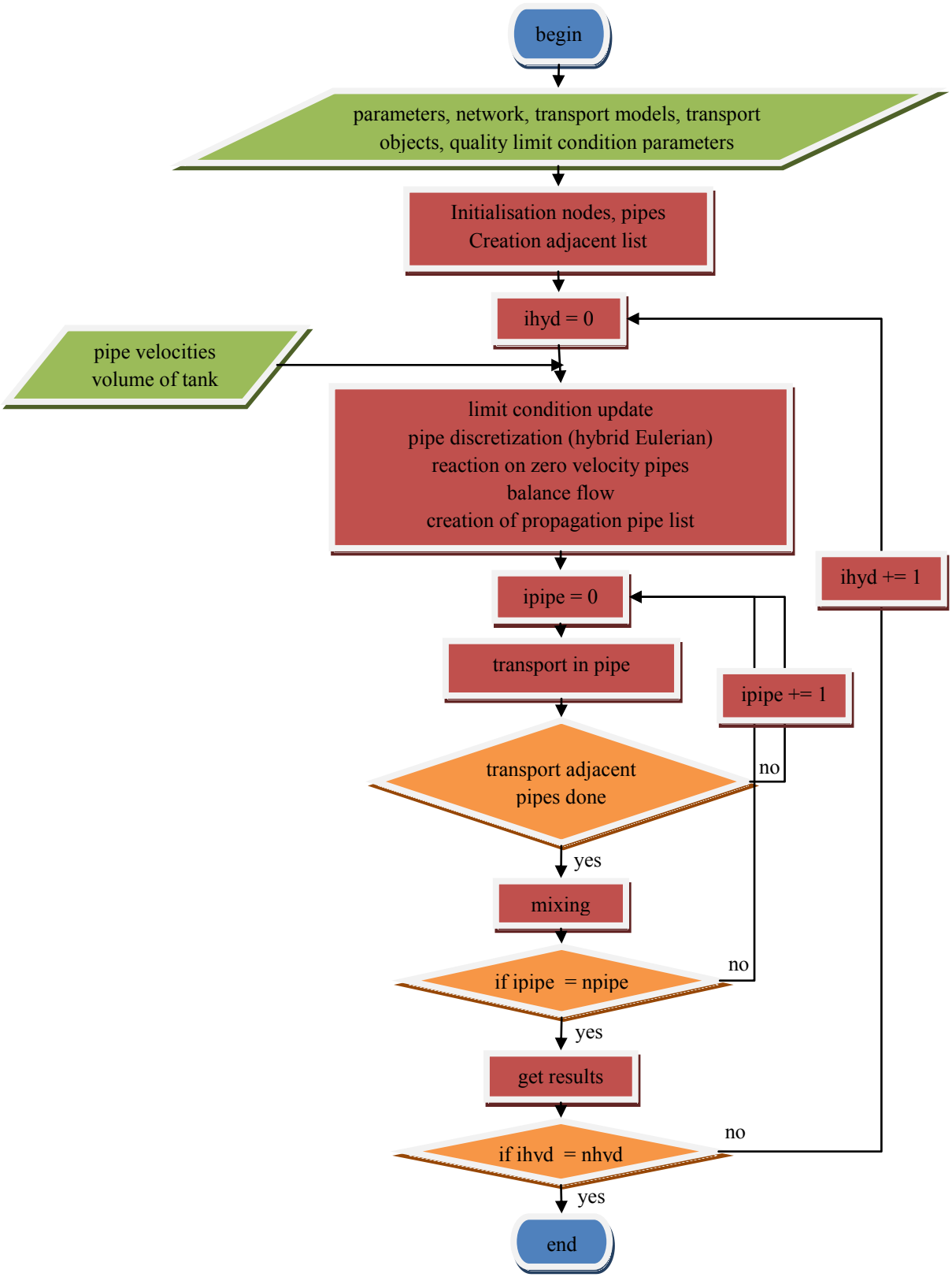


Figure 75: Transport algorithm.

A-V Sub Algorithms

It is described in this part all the sub algorithms that may be used to solve the transport model.

A-V-1 Pipe Discretization

Be u the velocity, L the length of the pipe, NbPint the previous interior number point, newNbPint the new one, NbPintVelZero a constant and value the previous quality results.

For all pipes:

if $(u \neq 0)$*:*

Calculate interior point number $newNbPint = ceil\left(\frac{L}{u}\right) - 1$

else:

$newNbPint = NbPintVelZero$

if $(newNbPint \neq NbPint)$*:*

Interpole value:

for i = 0 to $newNbPint$ *+1:*

newi = floor((NbPIint+1/ $newNbPint$ *+1)×i)*

position = NbPIint+1/ $newNbPint$ *+1×i-newi*

newvalue[i] = (value[newi+1]-value[newi])×position + value[newi]

A-V-2 Adjacent List

Be nodeadjacent the adjacent list, begnode and endnode the list of node at begin and end of pipe.

For all pipes ipipe:

Add ipipe to nodeadjacent[begnode[ipipe]]

Add ipipe to nodeadjacent[endnode[ipipe]]

A-V-3 Reaction

Be concentration the previous concentration, K and α the constant and order of the reaction and dt the time step of reaction.

if $\alpha = 0$*:*

newconcentration = concentration –K×dt

else if $\alpha = 1$*:*

$$newconcentration = concentration \times e^{-K*dt}$$

else

$$newconcentration = concentration - K \times concentration^{\alpha} \times dt$$

A-V-4 Temperature

Be temperature the previous temperature, K_T the thermal constant, T_0 the extern temperature and dt the time step.

$$newtemperature = temperature - K_T \times (temperature - T_0) \times dt$$

A-V-5 Hydraulic Balance

Be flowout and flowin the outflow and inflow list for every node, and endnode the list of node at begin and end of pipe and flow the flow value for each pipe.

For all pipes ipipe:

if flow[ipipe] > 0:

flowout[begnode[ipipe]] = flowout[begnode[ipipe]] + abs(flow[ipipe])

flowin[endnode[ipipe]] = flowin[begnode[ipipe]] + abs(flow[ipipe])

else

flowin[begnode[ipipe]] = flowout[begnode[ipipe]] + abs(flow[ipipe])

flowout[endnode[ipipe]] = flowin[begnode[ipipe]] + abs(flow[ipipe])

For all nodes:

if flowout[ipipe] > flowin[ipipe] + ε:

ipipe is inflow with flow = flowout[ipipe] - flowin[ipipe]

A-V-6 Propagation List

Be mark a list of boolean on pipe size,

For all pipes:

mark[i] = false

For all limit, initial condition and inflow nodes:

(R) For all pipe ipipe adjacent to node:

If mark[ipipe] = false:

mark[ipipe] = true

add ipipe to propagation list

recursivity on (R) with other pipe end node

A-V-7 Hybrid Transport

Be C the concentrationvalue table for one pipe, u the velocity, idx the spatial position, idt the time, and dt the time step:

$$dt = dt_{moc} + dt_{\theta}.$$

For all dx in pipe:

C[dx+1, idt+dt_{moc}] = C[dx, dt]

For all dx in pipe:

C[dx+1, idt+dt] = C[dx+1, idt+dt_{moc}] -h×(θ ×(C[dx+1, idt+dt_{moc}] - C[dx, idt+dt_{moc}])-(1- θ)× C[dx, idt+dt])

C[dx+1, idt+dt] = reaction(C[dx+1, idt+dt])

with

$$h = \frac{u * dt}{dx}.$$

A-V-8 Adjoint Transport

Be time the current time, length the current length, u the velocity, dt the time step, begnode and endnode the pipe ends and L the length of the pipe.

While time > 0

length = length – u×dt

if (length < 0)

node = begnode

$$time = time - \frac{length}{u}$$

end

else if (length > L)

node = endnode

$$time = time - \frac{L - length}{u}$$

end

else

time = time – dt

A-V-9 Average Mixing

Be CIC, CinjC and Cin_{ipipe} the concentration of initial condition, injection and pipe ipipe respectively. Be QIC, QinjC and Qin_{ipipe} the flows and Qinflow eventual flow going into the node.

$$\text{concentration} = \frac{QIC * CIC + QinjC * CinjC + \sum_{ipipe}^{pipes} Qin_{ipipe} Cin_{ipipe}}{QIC + QinjC + \sum_{ipipe}^{pipes} Qin_{ipipe} + Qinflow}.$$

Appendix B: Imperfect Mixing and Dispersion

B-I Imperfect Mixing

First the imperfect mixing function is given for one double T-junction. It gives the concentration at each outputs from the ones at inputs. Then is explained how it is adapted to current transport models.

B-I-1 Imperfect Mixing Algorithm

The aim of this function is to calculate the concentration at the beginning (node) of a pipe inside a double T-junction configuration. The following equation will be used:

$$C_{out} = \frac{q_{in1}C_{in1}a_{out} + q_{in2}C_{in2}(1 - a_{out})}{q_{out}},$$

with q the flow, C the concentration and a (between 0 and 1) the percentage of concentration for the output coming from input 1. It respects the following conditions:

$$\begin{cases} a_{out,in2} = (1 - a_{out,in1}) \\ a_{out1} = (1 - a_{out2}) \end{cases}.$$

Inputs/outputs:

Inputs:

-distance L (double in meter)

-Reynolds number Re1,Re2,Re3,Re4 (double)

-diameter D1,D2,D3,D4 (double in meter)

-concentration C1,C2 (at the end of the input pipes)

-boolout (boolean: 0 if 1st output, 1 if 2nd output)

Output:

-concentration at the beginning of the output pipe (1st or 2nd depending on boolout)

The data table is given under the form [length (in diameter), Reynolds number, ReIn1, ReIn2, ReOut1, ReOut2, mass flow percentage at Out1]

This table is proceed into another table under the form [length (in diameter), Reynolds number, %In1, %Out1, mass flow percentage at Out1-%Out1]

Then, previous to calculation the inverse of the covariance matrix invCov is calculated. Q1, Q2, Q3 and Q4 are calculated from Reynolds numbers and diameters.

If 5D < L < 20D and Reynolds number > 1,000

Cov0 = covariance between point of calculation and data

B = invCov×Cov0

θ = B×data_results

coeff = (θ + %Out1) /100

if coeff > 1 then coeff = 1and if coeff < 0 then coeff = 0

coeff2 = (Q3-coeff×Q1)/Q2;

if coeff2 > 1 then coeff2 = 1 and if coeff2 < 0 then coeff2 = 0

$$C3 = \frac{coeff2 \times Q1C1 + coeff \times Q2C2}{Q3}$$

$$C4 = \frac{(1 - coeff2) \times Q1C1 + (1 - coeff) \times Q2C2}{Q4}$$

Else

$$C3 = C4 = \frac{Q1C1 + Q2C2}{Q1 + Q2}$$

B-I-2 Hybrid Eulerian Adaptation

When doing transport, test if the pipe belongs to a double T-junction, then change first discretization value accordingly.

B-I-3 Forward Adjoint Adaptation

During mixing, take correct value if one inflow pipe belongs to a double T-junction.

B-II Dispersion

The dispersion algorithm is an improved version of the adjoint transport equation applied to graph as explained chapter II. It includes both particle backtracking and random walk to model both velocity profile and radial diffusion effect.

B-II-1 Main Algorithm

Be Np the number of particle, N number of discretization on the radius, u the velocity, dt the time step, x the axial position, r the radius, R the radius of the pipe, D_m the diffusive coefficient, time the current time, concentration the concentration list values.

For all adjacent pipe:

if u>0:

newconcentration = 0

$$dr = \frac{R}{N}$$

for all N discretization idr:

$$newconcentration(ir) = 0$$

for all Np particles:

While (particle still inside the pipe and time > 0):

x = x – u(r)×dt

r = random_walk(r, D)

time -= dt

$$newconcentration(ir) += concentration\ [time]$$

$$newconcentration(ir) = \frac{newconcentration(ir)}{Np}$$

$$drintegrale = \frac{(r+\frac{dr}{2})^2-(r-\frac{dr}{2})^2}{R^2}$$

$$newconcentration += drintegrale * newconcentration(ir)$$

B-II-2 Random Walk

Be dt the time step, r the radius, R the radius of the pipe and D_m the diffusive coefficient.

dx = Gaussian(D_m ×dt)

dy = Gaussian(D_m ×dt)

$newr = \sqrt{(r + dx)^2 + dy^2}$

if newr < R:

r = newr

else

do the symmetry on the circle boundary

B-II-3 Gaussian Random Function

Be $\eta_1$and $\eta_2$two random values between 0 and 1.

results = $\sqrt{2D_m dt}\sqrt{(-2 * ln(\eta_1)}cos\ (2\pi \times \eta_2)$

Appendix C: Source Identification and Sensor Placement

C-I Source Identification

The source identification method includes both sensor response simulation, creation of the input/output contamination matrix and the scoring.

C-I-1 Sensor Responses

Either define a contamination scenario: the number of contamination with the node of the contamination, the begin time and the duration of each. Or have N random single contamination scenario.

Then give the transport model and its parameters. The forward adjoint model needs the parameter Np the number of particle, N number of discretizations on the radius and D_m the diffusive coefficient.

For all simulation, the node of contamination gives an injection (needed for multi contamination) value of either 0 or 1.

The contamination results at sensors are used as contamination responses

$$\begin{cases} \text{false if concentration} < threshold \\ \text{true else} \end{cases}$$

with threshold a constant between 0 and 1.

C-I-2 Recursive Backtracking on Graph

The backtracking algorithm or adjoint transport method (see A-V-8) is used for source identification. From one node at certain time, it traces back a particle to know where and when that particle entered any inflow pipe.

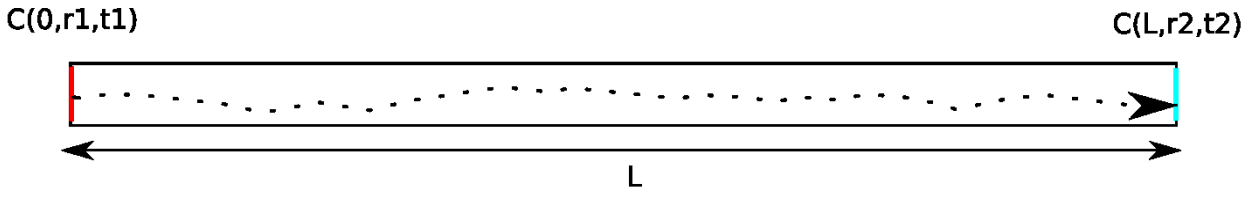


Figure 76: Backtracking.

By recursivity that function can be extended to the whole graph and trace back the potential source of it.

It can be used in different subjects. It can help in risk analysis by giving risk for a node depending on its potential sources. It can also serve for sensor placement by calculating the detection coverage of a node. Finally it is used hereafter for source identification with the construction of the imput/output matrix.

Be a node inode and time itime

(R) If sensor response is positive:

Add (inode, itime) to source list

(inewnode, inewtime) = backtracking(inode, itime) (see A-V-8)

do (R) with (inewnode, inewtime)

C-I-3 Source Identification Algorithm

For all itime:

inode = isensor

(Rinv) If sensor response is negative:

Matrice[itime, inode] = -1

(inewnode, inewtime) = backtracking(inode, itime) (see A-V-8)

do (Rinv) with (inewnode, inewtime)

(R) If sensor response is positive:

If (Matrice[itime, inode] ≠ -1)

Add (inode, itime) to input output matrix

(inewnode, inewtime) = backtracking(inode, itime) (see A-V-8)

do (R) with (inewnode, inewtime)

For all potential node inode:

score[inode] = 0

If (Matrice[itime, inode] ≠ -1):

for all positive responses:

if coeff = 1:

score[inode] += 1

$$score\ [inode]\ =\ \frac{score[inode]}{positive\ response\ number}$$

Rank nodes upon score.

C-II Sensor Placement

The sensor placement problem is defined in chapter V. All resolution methods presented here use a Monte Carlo algorithm as base to get for instance average time to detection, detection likelihood or Contribution to source identification.

C-II-4 Monte Carlo

The Monte Carlo method is a statistical method based here on contamination scenario.

First, a certain number are randomly selected at a node, at a time of injection and for a certain duration. Then the transport is simulated and the results save in a table. That table is then proceeded to extract the information wanted.

In both the greedy and local search algorithm proposed hereafter the objective function is the Contribution of the real source of contamination. It is calculated for each scenario with the help of the source identification algorithm presented before.

C-II-5 Greedy Algorithm

Be N the number of sensor expected to be placed on the network, a list of node detailing the different nodes that may be used as sensor location and a list of contamination scenarios.

Begins with 0 sensor.

While number sensor < N:

Loop on the list of potential sensor:

Loop on contamination scenario:

Get Contribution

Add the sensor with the highest Contribution

C-II-6 Local Search

Be an initial sensor placement such as the one given by Piller (2015) with average time to detection as objective function.

Begins with a N initial sensor placement

While find a higher Contribution:

Loop on the list of sensor isens:

Loop on adjoint node to isens:

get Contribution when adjoint node is taken in place of isens

Loop on the list of sensor isens:

Replace isens to the adjoint node with max Contribution if it is higher than isens Contribution

Appendix D: Object model

All algorithms are gathered in one object model:

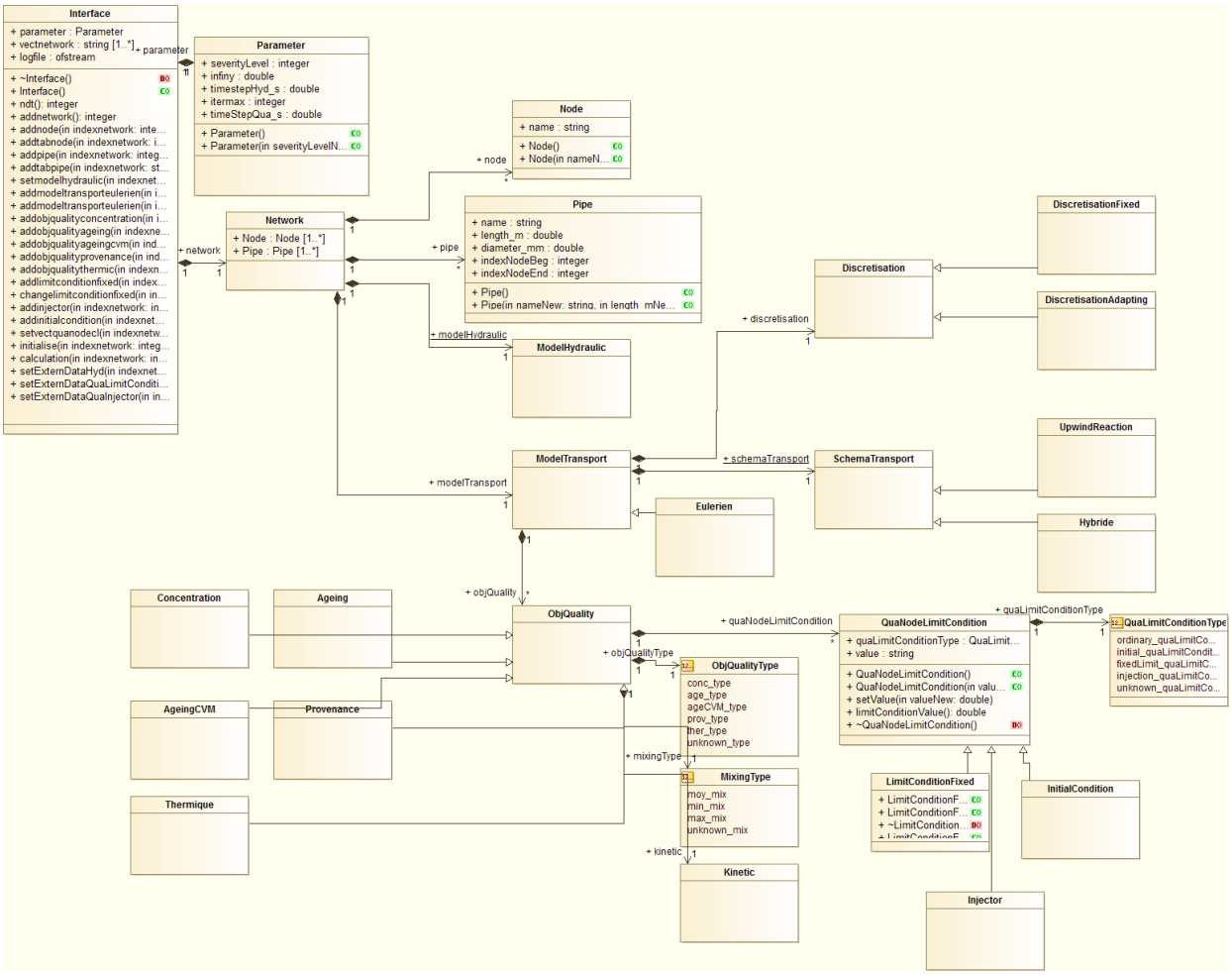


Figure 77: Object model.

Unified Modeling Language has been used to create a coherent domain for all algorithms to interact between them. It has already been proved that object modeling help the development of code and its sustain.

It helps mostly the development of new functions by focusing only on parts of the code whose links to other function are well defined. For instance an upwind scheme can easily replace the hybrid transport model defined as an Eulerian transport scheme. Therefore new functionalities can both be implemented and tested. This is also particularly interesting in the research field as new works come complete precedent ones and algorithm needs to be tested and compared to previous ones.

Also different versions of same object can be implemented to differ only on some parameters. For example Concentration and Ageing objects differ here only on the fixed parameter K and α, constant and order of reaction, whose values are -1 and 0 respectively. That is interesting to propose formatted version of quality object to users.

The object model sustains the code development. Indeed, classic code may contain repetition of code applied to different objects. However, object modeling allows the gathering of different object into one category upon which same function is applied on. Therefore change of that function will affect all of them.

Another important use is inheritance of attributes and methods. Both can be initialized differently as well as behave differently depending on the object they belong to. For instance, Eulerian model and forward adjoint model both have a transport function. Both are called in the same way by the main function, however their implementations are completely different. One used Eulerian transport on pipes, the other one use adjoint transport from node to node.

UML not only allows the user to create the class diagram whose use is to define the objects and create the relation between them. Different level of conception can be define, from the user case to the technical point of view.

Object model allows also code development managing, tools such as Modelio that have been used here, allows several users to work together on the same project. One may use multi user lock option to forbid other user modifying the files they are working on. Versioning can also be done to do automatic savings as well as managing the merging process.

Finally automatic documentations can be done through object modeling. Any commentaries can both be process directly to generate documentation but also commentaries inside the code. That eases the coding process as well as gives a better representation of the whole program.